

**TRABAJO FIN DE MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA
DE NEGOCIOS**

**USO DE TÉCNICAS DE WEB MINING: APLICACIÓN EMPÍRICA
EN EL SECTOR DE LA ADMINISTRACIÓN PÚBLICA**

**FACULTAD DE ESTUDIOS ESTADÍSTICOS
UNIVERSIDAD COMPLUTENSE DE MADRID**



ALUMNO: JORGE EUSEBIO VELASCO LÓPEZ

TUTORA: MARIA JOSÉ NARROS

CURSO ACADÉMICO: 2012/2013

CONTENIDO

CONCEPTUALIZACIÓN

I. INTRODUCCIÓN	3
APROXIMACIÓN A LA MINERÍA DE DATOS WEB (<i>WEB MINING</i>)	3
TAXONOMÍA DE LA WEB MINING	7
WEB USAGE MINING	8
JUSTIFICACIÓN DE LA NECESIDAD DE LA INVESTIGACIÓN	11

OBJETIVOS Y METODOLOGÍA

II. OBJETIVOS Y CALENDARIO	12
III.-ANÁLISIS FUNCIONAL DEL SISTEMA DE INFORMACIÓN	14
CONSTRUCCIÓN DEL MODELO DE PROCESOS DEL NUEVO SISTEMA	14
UNIDADES AFECTADAS, SUS FUNCIONES E INTERRELACIÓN	14
IDENTIFICAR LA OPCIÓN TECNOLÓGICA ELEGIDA	14
IDENTIFICACIÓN Y DEFINICIÓN DE SUBSISTEMAS	15
CONSTRUCCIÓN DEL MODELO DE DATOS DEL NUEVO SISTEMA.	24
ARQUITECTURA DE LA INFORMACION	24
ESQUEMA FÍSICO DE LOS DATOS.	25
DESCRIPCIÓN DE LOS DATOS	26

DESARROLLO DE LA INVESTIGACIÓN Y CONCLUSIONES

IV.- ANÁLISIS Y EJECUCIÓN DEL PROCEDIMIENTO DE MINERÍA WEB.....	30
ANÁLISIS EXPLORATORIO DE DATOS	30
CONSTRUCCIÓN DEL MODELO.....	40
COMPARACIÓN DE LOS MODELOS	57
V.- SUMARIO DEL PROCEDIMIENTO Y RESÚMEN DE RESULTADOS	59

ANEXOS

VI.- ANEXO SOBRE DETALLE DEL WEB CONTENT Y WEB STRUCTURE MINING	62
VII.- ANEXO SOBRE NOCIONES BÁSICAS DE TECNOLOGÍA WEB.....	66
VIII.- ANEXO DE BIBLIOGRAFIA	71

I. INTRODUCCIÓN

APROXIMACIÓN A LA MINERÍA DE DATOS WEB (*WEB MINING*)

El crecimiento de la información que se encuentra en la Web ha sido desmedido debido a la necesidad de los usuarios (personas físicas, empresas, universidades, Gobierno, etc.) de contar con datos para la interrelación en el mundo globalizado. Según Baeza-Yates [consultar el Anexo de Bibliografía, 1], la información de la Web es finita pero el número de páginas web es infinita.

Internet es una especie de universo paralelo digital que, de manera análoga al que conocemos, se expande y se contrae. Cada día en la *World Wide Web* se crean y se destruyen millones de páginas web, y ahora, gracias a “www.worldwidewebwize.com” podemos saber cuántas páginas web existen en Internet en cada preciso instante.

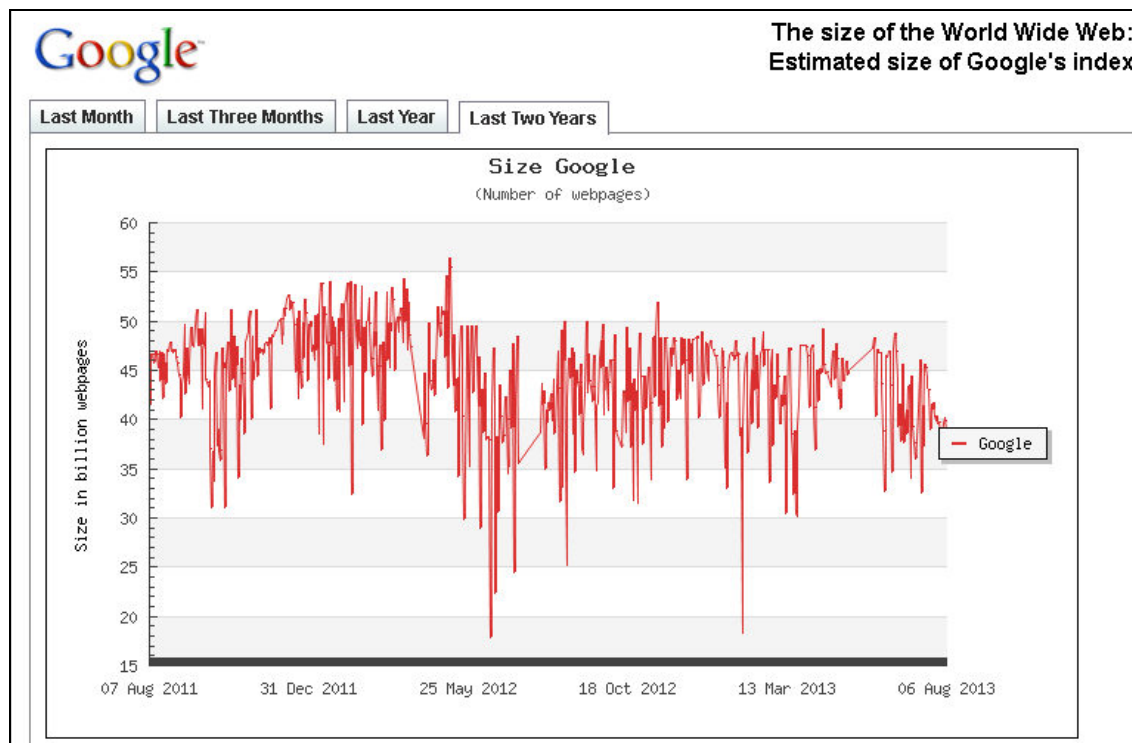


Imagen 1.1. Tamaño estimado de la *World Wide Web*, según la indexación de Google.

Actualmente existen alrededor de 40 mil millones de páginas indexables. Es decir, la información que poseen los buscadores Web. Sin embargo, es importante mencionar que la mayoría de las páginas web no son indexables ya que son dinámicas, es decir son aquellas que se generan automáticamente con datos extraídos de bases de datos [2].

Existen diferentes problemas a los que se enfrentan los usuarios debido a este crecimiento exponencial. Uno es el que representa encontrar información relevante debido a dos aspectos fundamentales, la baja precisión y la escasa cobertura. Esta última es porque no todos los motores de búsqueda tienen la suficiente capacidad de indexar la Web, debido a varios factores: el ancho de banda, el espacio de disco duro, el coste económico, etc.

La *Web Mining* actualmente es un área de investigación extensa dentro de varios grupos de trabajo, especialmente interesados debido al alto crecimiento de la información que existe en la Web y por el movimiento económico que ha generado el *e-commerce* y sobre todo para intentar resolver los problemas que se han mencionado anteriormente, ya sea de manera directa o indirecta. Actualmente, el objetivo principal es aprender de comportamientos de los usuarios en su andar por la Web y así proporcionarles información realmente relevante, útil y personalizada en muchos casos.

La presente introducción pretende adentrarnos un poco en el mundo de la *Web Mining*, permitiéndonos conocer los aspectos básicos tales como, el proceso general de la *Web Mining* así como unas breves aproximaciones de la *Web Content Mining* (de contenido), *Web Structure Mining* (de estructura) y de la *Web Usage Mining* (de uso).

PARTICULARIDADES DE LA WEB

Es común encontrarse en la Web con ciertas características preponderantes que pudieran parecer problemas. Sin embargo, podrían catalogarse como oportunidades sin precedentes para la obtención de información y mejora de la Web. Por ejemplo, la información en la Web se presenta en cantidades descomunales, fácilmente accesibles y por lo general, heterogénea. Es decir, muchas páginas presentan la misma o similar información usando formatos diferentes.

La Web normalmente está compuesta por una mezcla de tipos de información. Por ejemplo, contenido principal, anuncios, paneles de navegación, noticias de *copyright*, etc. Para una aplicación en particular sólo parte de la información es útil y el resto es basura o inútil [3].

La idea de Minería de Datos no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como *Data Fishing*, *Data Mining (DM)* o *Data Archaeology* con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido.

A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro entre otros, empezaron a consolidar los términos de Minería de Datos y Knowledge Discovery and Data Mining (KDD).

Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

La evolución de sus herramientas en el transcurso del tiempo puede dividirse en cuatro etapas principales: 1.- Recolección de Datos (1960); 2.- Acceso de Datos (1980) ; 3.- Almacén de Datos y Apoyo a las Decisiones (principios de la década de 1990); 4.- Minería de Datos Inteligente (finales de la década de 1990).

Algunos autores definen a la *Web Mining* como el uso de técnicas para descubrir y extraer de forma automática información de los documentos y servicios de la Web.

Según M. Scotto [4] *Web Mining* es el proceso de descubrir y analizar información “útil” de los documentos de la Web. Sin embargo y teniendo en cuenta lo expuesto en la introducción, la minería Web se puede definir como el descubrimiento y análisis de información relevante que involucra el uso de técnicas y acercamientos basados en la minería de datos (*Data Mining*) orientados al **descubrimiento y extracción automática de información de documentos y servicios de la Web, tomando en consideración el comportamiento y preferencias del usuario.**

En la *Web Mining*, los datos pueden ser coleccionados en diferente niveles; en el área del servidor, en el lado del cliente (*cookies*), en los servidores *proxys (log files)*, etc.

De acuerdo con Etzioni [5] el proceso general de la *Web Mining* es el siguiente:

1. **Recuperación de Información:** Sería básicamente el proceso del descubrimiento automático de documentos relevantes de acuerdo a una cierta búsqueda. Documentos relevantes disponibles en la Web tales como noticias electrónicas, *newsgroups*, *newswires*, contenido de las html, etc. así como *logfiles* o *cookies*.

2. **Extracción de Información:** Tiene como objetivo transformar la información extraída en el proceso de recuperación de información, en documentos más fáciles de leer y de analizar.
3. **Generalización:** Reconocimiento de Patrones generales de una página en particular o bien también patrones de diferentes páginas.
4. **Análisis:** Una vez que los patrones han sido identificados, la parte humana juega un papel importante haciendo uso de herramientas adecuadas para entender, visualizar e interpretar los patrones

MOTORES DE BÚSQUEDA

Los motores de búsqueda han tenido un éxito considerable gracias a la enorme necesidad que tienen los usuarios de indagar en información no disponible en sitios físicos pero sí en Internet. Un motor de búsqueda tiene como objetivo indexar archivos almacenados en los servidores Web, por ejemplo los buscadores de Internet.

El resultado de la búsqueda es un listado de direcciones Web en los que se mencionan temas relacionados con las palabras clave buscadas [6].

Basándonos en las descripciones realizadas por Henzinger [7], podemos decir que los motores de búsqueda están integrados básicamente por tres grandes componentes:

1. *Crawler*: Recorren las páginas recopilando información sobre sus contenidos. Recolectan páginas de manera recurrente a partir de un conjunto de Links de páginas iniciales. Cuando buscamos una información en los motores, ellos consultan su base de datos, y nos la presentan clasificada por su relevancia. Los buscadores pueden almacenar cualquier tipo de página.
2. Indexador: el objetivo principal del indexador es procesar las páginas recopiladas por el *crawler*. La indexación proporciona agilidad en las búsquedas, lo que se traduce en mayor rapidez a la hora de mostrar resultados.
3. *Query processor*: procesa las consultas de los usuarios y devuelve resultados de acuerdo a esas consultas y de acuerdo a un algoritmo de posicionamiento.

Estos *crawlers* (o *spiders* en adelante) deben ser identificados y aislados en un análisis de minería Web, puesto que no son el objetivo del análisis.

Minería de contenido web

Su objetivo es la recogida de datos e identificación de patrones relativos a los contenidos del *Website* y a las búsquedas que se realizan sobre los mismos. Es decir son los datos reales que se entregan a los usuarios, los datos que almacenan los sitios Web [8].

La minería de contenidos, consiste en el análisis de datos desestructurados tales como texto libre, semi-estructurado como documentos HTML y más estructurados como datos en tablas o páginas generadas con datos de Bases de Datos.

Existen dos grupos de estrategias sobre minería de contenidos: aquellas que “minan” directamente el contenido de los documentos y aquellas que mejoran en la búsqueda de contenidos.

Minería de estructura web

La minería de estructura intenta descubrir el modelo subyacente de las estructuras de los enlaces del Sitio web (*Website*). El modelo se basa en la topología de los hiperenlaces con o sin la descripción de los enlaces. Este modelo puede ser usado para categorizar las páginas web y es útil para generar información tal como la similitud y relación entre diferentes páginas web [9]. Es decir, pretende revelar la estructura real de un sitio Web a través de la recogida de datos referentes a su estructura y, principalmente a su grado de conectividad.

Minería de uso Web

La minería de uso intenta dar sentido a los datos y comportamientos generados en las sesiones de navegación del *Website*. Es decir, son aquellos datos que describen el uso al cual se ve sometido un sitio Web, registrado en los logs de acceso a de los servidores Web.

A partir de esta información se podría concluir, por ejemplo, qué documento visitado no tiene razón de ser, o si una página no se encuentra en los primeros niveles de jerarquía de un sitio [10]. Analizar los logs de diferentes servidores Web, puede ayudar a entender el comportamiento del usuario, la estructura del *Website*, permitiendo de este modo mejorar el diseño de esta colección de recursos.

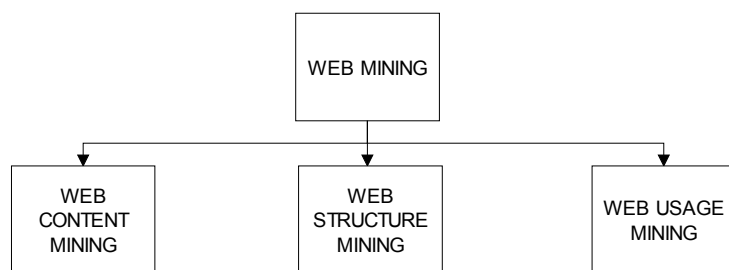


Figura 1.1. Taxonomía de la *Web Mining*

En el Anexo se puede consultar más detalle sobre la minería de contenido y estructura Web, ahora nos centramos en el que es objeto del presente estudio.

WEB USAGE MINING

Los logs que se generan constantemente en los servidores debido a los requerimientos de los usuarios, generan un gran volumen de datos provenientes de dichas acciones. Recientemente este gran volumen de información relevante ha empezado a usarse para obtener **datos estadísticos**, analizar **accesos inválidos** y para **analizar problemas que se produjeran en el servidor**.

Los datos almacenados en los logs siguen un formato estándar. Una entrada en el log siguiendo este formato contiene entre otras cosas, lo siguiente: dirección IP del cliente, identificación del usuario, fecha y hora de acceso, requerimiento, URL de la página accedida, el protocolo utilizado para la transmisión de los datos, un código de error, agente que realizó el requerimiento y el número de bytes transmitidos, siendo cada acceso es un registro distinto (consultar Anexo para detalle).

Association Rules, *Sequential Patterns* y *Clustering* ó Clasificación son algunas de las técnicas de *data Mining* que se aplican en los servidores Web.

ASSOCIATION RULES

Las técnicas de Reglas de Asociación o *Association Rules* juega un papel muy importante en el contexto de la nueva visión de la Web. Es decir, con el auge de las técnicas de comercio que se manejan de forma electrónica permiten el desarrollo de estrategias voraces de marketing.

Normalmente esta técnica está relacionada con el uso de Bases de Datos Transaccionales, donde cada transacción consiste en un conjunto de ítems. En este modelo, el problema consiste en

descubrir todas las asociaciones y correlaciones de ítems de datos donde la presencia de un conjunto de ítems en una transacción implica la presencia de otros ítems.

Esta técnica generalmente está asociada con el número de ocurrencias de los ítems dentro del log de transacciones [11]. Por lo tanto, podemos identificar la cantidad de usuarios que acceden a determinadas páginas (un ejemplo hipotético podría ser que del 60% de los clientes que acceden a la página con URL `"/ine/productos/"`, también acceden a la página `"/ine/productos/producto1.html"`).

Por otro lado nos permite mejorar considerablemente la estructura de nuestro *site*, por ejemplo, si descubrimos que el 80% de los clientes que acceden a `"/ine/productos"` y `"/ine/productos/file1.html"`, también acceden a `"/ine/productos/file2.html"`, parece indicar que alguna información de `"file1.html"` lleva a los clientes a acceder a `"file2.html"`.

Esta correlación podría sugerir que esta información debería ser movida a `"/ine/productos"` para aumentar el acceso a `"file2.html"`.

SEQUENTIAL PATTERNS

En general, en las Bases de Datos transaccionales se tienen disponibles los datos en un período de tiempo y se cuenta con la fecha en que se realizó la transacción; la técnica de Patrones de Secuencias o *Sequential Patterns* se basa en descubrir patrones en los cuales la presencia de un conjunto de ítems es seguido por otro ítem en orden temporal.

En el log de transacciones de los servidores de un *Website*, se guarda la fecha y hora en la que un determinado usuario realizó los requerimientos. Analizando estos datos, se puede determinar el comportamiento de los usuarios con respecto al tiempo.

Aunque este análisis transversal en el tiempo no se ha hecho en el presente estudio, se podría determinar hipotéticamente, que el 60% de los clientes que accedieron a `"/ine/productos/producto1.html"`, también accedieron a `"/ine/productos/producto4.html"` dentro de los siguientes 7 días.

El descubrimiento de *sequential patterns* en el log puede ser utilizado para predecir las futuras visitas y así poder organizar mejor los accesos para determinados períodos. Por ejemplo, utilizando esta técnica, se podría descubrir que los días laborables entre las 9 y las 12 horas

muchas de las personas que accedieron al servidor lo hicieron para ver la nota de prensa, pudiéndose facilitar su acceso durante esa franja de tiempo.

También puede ser utilizado para descubrir tendencias, comportamiento de usuarios, secuencias de eventos, etc. Esta información puede ser aprovechada tanto en el aspecto comercial (CRM) como en el aspecto técnico (mejorar los tiempos de acceso).

En general, todas las herramientas que realizan Mining sobre el log enfocan el análisis sobre secuencias de tiempo ya que los eventos que son almacenados están muy relacionados con el ámbito temporal en que se producen.

CLUSTERING

Las técnicas de clasificación permiten desarrollar un perfil para los ítems pertenecientes a un grupo particular de acuerdo con su perfil multivariable. Este perfil posteriormente puede ser utilizado para clasificar nuevos ítems que se agreguen en la base de datos.

En el contexto de la *Web Mining*, las técnicas de clasificación permiten desarrollar un perfil para clientes que acceden a páginas o archivos particulares, basado en la información demográfica disponible de los mismos. Esta información puede ser obtenida analizando los requerimientos de los clientes y la información transmitida de los *browsers* incluyendo el URL.

La información acerca de los clientes puede ser obtenida del *browser* del cliente automáticamente por el servidor; esto incluye los accesos históricos a páginas, el archivo de *cookies*, etc. Otra manera de obtener información es por medio de los formularios *online* de registro, aunque esta vía no estaría integrada por el momento en el INE.

La agrupación automática de clientes o datos con características similares sin tener una clasificación predefinida es llamada *clustering*.

Se puede obtener una aproximación a fundamentos de terminología Web y a la estructura del fichero log a los que se aludirán en el presente documento, en el Anexo.

JUSTIFICACIÓN DE LA NECESIDAD DE LA INVESTIGACIÓN

La *Web Mining* es un área con espectro amplio de investigación. Los Sistemas basados en la Web son cada vez más, la tecnología más utilizada para la divulgación de los productos y servicios en toda Organización y en particular, para la difusión de productos estadísticos contemplados por el Instituto Nacional de Estadística (INE).

Esto es así debido a la facilidad de utilización y disponibilidad de las herramientas para navegar por la Web así como de la facilidad en el desarrollo y mantenimiento de los recursos Web. Su desarrollo ha tenido un crecimiento constante durante los últimos años y esto también ha motivado la aplicación de técnicas de minería de datos o descubrimiento de conocimiento que ya se han aplicado con éxito en sistemas de comercio electrónico o *e-commerce* para comprender el comportamiento de clientes en línea con la finalidad de poder incrementar las ventas.

Estas herramientas inteligentes utilizan técnicas de extracción de conocimiento para descubrir información útil y así poder mejorar el sistema.

La aplicación de técnicas de minería de datos se puede contemplar desde dos perspectivas complementarias:

1. **Desde la perspectiva del equipo de Difusión/TIC.** Con el objetivo de obtener conocimiento para mejorar el funcionamiento o rendimiento de sus Sistemas a partir de la información sobre la utilización de la Web por parte de los usuarios. Las principales aplicaciones serían sistemas de personalización, sistemas recomendadores, sistemas de detección de irregularidades..., debido a sus capacidades para el descubrimiento de patrones de navegación regulares e irregulares. También clasificaciones de usuarios y de los contenidos, construcción adaptativa de contenidos, descubrimiento de relaciones entre actividades, etc.
2. **Desde la perspectiva de los usuarios.** Con el objetivo de mejorar la experiencia de usuario. Sus principales aplicaciones serían: sugerir buenas experiencias de uso a los usuarios, adaptación de los contenidos según el *expertise* del usuario, ayudar a los usuarios dando sugerencias y atajos personalizados, etc.

Hay que señalar el importante esfuerzo por parte de la Administración Pública y por el INE en particular, por modernizar sus sistemas e implementar un servicio de calidad acorde con las

nuevas tecnologías y las necesidades de los usuarios. Este estudio estaría enfocado a ser uno más de dichos esfuerzos.

En la actualidad, se utilizan sistemas exploratorios de los *Weblogs* (*software* Analogs) y a partir del uso de *cookies* como el Google Analytics (GA). Ambos aportan resultados distintos y además de ámbito descriptivo y poco personalizable.

Uno de los objetivos será por tanto contrastar los resultados de esos sistemas exploratorios a partir de los *Weblogs*. Después se realizará un análisis de uso aplicando las técnicas mencionadas en el punto anterior y se desarrollará un aplicativo para poder realizar minería de Web en el futuro.

II. OBJETIVOS Y CALENDARIO








El **objetivo principal** de este trabajo por tanto, es la investigación, aplicación y combinación de técnicas de clasificación y reconocimiento de patrones, para procesar la información disponible sobre el comportamiento de usuarios, en el entorno de la Web en la administración pública y en particular del INE.

Objetivos Específicos

1. Realizar un pre-procesamiento de la información recopilada y comparar resultados descriptivos de *softwares* en uso.
2. Aplicación y combinación de metodologías de reconocimiento de patrones para la obtención de tipologías y perfiles de usuarios para la generación de conocimiento para la Organización.
3. Uso de otras técnicas estadísticas (cadenas de Markov, SEP, etc.) para generación de conocimiento del modelo de uso de la Web del INE por parte de los usuarios.
4. Desarrollar un sistema de información que permita la obtención de una herramienta en lenguaje SAS/AF para preprocesar los datos de los logs y aplicar automáticamente técnicas de reconocimiento de patrones a un primer nivel.

Para la parte informática del presente proyecto, Webclicks INE, se va a realizar un análisis funcional abreviado del Sistema, como si se tratara de un nuevo Sistema de Información a crear. El análisis de *Web Mining* se expondrá una vez se analice el Sistema de Información.

El calendario de desarrollo sería el esquematizado a continuación:

1		Weblogs INE	54,875 days?	24/06/13 9:00	6/09/13 17:00	
2		Inicio del Proyecto	6,75 days?	24/06/13 9:00	2/07/13 16:00	
3		Elaboración del Plan Director del Proyecto	5,875 days 4	24/06/13 9:00	1/07/13 17:00	
4		Reuniones con Sistemas y Servicio Promotor para Lanzamiento	0,875 days?	2/07/13 8:00	2/07/13 16:00	3
5		Arquitectura y Análisis	15,75 days?	24/06/13 9:00	15/07/13 16:00	
6		Diseño Técnico	14,875 days?	24/06/13 9:00	12/07/13 17:00	
7		Diseño Detallado de los Componentes	14,875 days?	24/06/13 9:00	12/07/13 17:00	
8		Fase 0: Exploración del web del INE	5,875 days? 9	24/06/13 9:00	1/07/13 17:00	
9		Fase 1. Obtención de los ficheros de Secuencias, Page, Weblogs41 y Weblogs44	4 days? 10	2/07/13 8:00	5/07/13 17:00	8
10		Fase 2. Programación aplicativo de obtención de Ficheros para Análisis y Obtención de F	5 days? 11	8/07/13 8:00	12/07/13 17:00	9
11		Aprobación Diseño Técnico	0,875 days? 40...	15/07/13 8:00	15/07/13 16:00	10
12		Diseño de Pruebas	54,625 days?	24/06/13 9:00	6/09/13 15:00	
38		Desarrollo y Prueba Unitaria	17,125 days?	15/07/13 16:00	7/08/13 17:00	
39		Desarrollo y Prueba Unitaria	17,125 days?	15/07/13 16:00	7/08/13 17:00	
40		Fase 0: Exploración del web del INE	2,125 days 42	15/07/13 16:00	17/07/13 17:00	11
41		Desarrollo y Prueba unitaria de la Funcionalidad	2,125 days	15/07/13 16:00	17/07/13 17:00	
42		Fase 1. Obtención de los ficheros de Secuencias, Page, Weblogs41 y Weblogs	8 days? 44...	18/07/13 8:00	29/07/13 17:00	40
43		Desarrollo y Prueba unitaria de la Funcionalidad	8 days?	18/07/13 8:00	29/07/13 17:00	
44		Fase 2. Programación aplicativo de obtención de Ficheros para Análisis y Obt	7 days? 46	30/07/13 8:00	7/08/13 17:00	42
45		Desarrollo y Prueba unitaria de la Funcionalidad	7 days?	30/07/13 8:00	7/08/13 17:00	
46		Pruebas	17 days?	8/08/13 8:00	30/08/13 17:00	44
47		Ejecución Pruebas Funcionales de Integración y Sistema.	8 days?	8/08/13 8:00	19/08/13 17:00	
54		Ejecución Pruebas de Rendimiento	9 days	20/08/13 8:00	30/08/13 17:00	52
55		Creación de Escenarios de Prueba	2 days	20/08/13 8:00	21/08/13 17:00	
59		Ejecución de Escenarios (x2)	1 day	23/08/13 8:00	23/08/13 17:00	
62		Gestión de Defectos	5 days	26/08/13 8:00	30/08/13 17:00	
65		Creación de Documentación y Análisis Web	30 days?	29/07/13 8:00	6/09/13 17:00	
66		Informes de Ejecución	30 days	29/07/13 8:00	6/09/13 17:00	11
67		Análisis web mining	29 days?	30/07/13 8:00	6/09/13 17:00	42

Imágen 2.1. Impresión de pantalla con el calendario de desarrollo

CONSTRUCCIÓN DEL MODELO DE PROCESOS DEL NUEVO SISTEMA

UNIDADES AFECTADAS, SUS FUNCIONES E INTERRELACIÓN

Las unidades afectadas por el Sistema Estadístico para WEBCLICKS INE y sus funciones a lo largo del ciclo de vida del sistema quedarían como se verá a continuación.

No hay unidades externas de entrada y de salida implicadas en el Sistema.

- **Entidades internas (Unidades Afectadas)**

- Sistemas: consolida los accesos de los log de los servidores Web del INE. Proporcionan análisis realizados sobre dichos logs a partir de software como el Analog.
- Difusión: analiza los informes proporcionados por Sistemas. Además dispone de otra herramienta de análisis a partir de otra tecnología, como es el Google Analytics.
- Desarrollo: realiza el desarrollo de la aplicación y de la metodología de análisis de los *Weblogs*.

El propio servicio promotor realizaría la tarea de Explotación del Sistema.

IDENTIFICAR LA OPCIÓN TECNOLÓGICA ELEGIDA

Para la realización del nuevo Sistema se va a utilizar el siguiente entorno:

- Sistema operativo Windows XP 32 bits.
- Ficheros planos para los logs.
- Programación en SAS 9.1 (SAS BASE, SAS STAT)
- Análisis de los datos en SAS Miner 9.2.
- Programación del aplicativo de preparación de los datos de entrada para análisis a partir de SAS/AF 9.1.
- En la Fase 0 como complemento metodológico se usará
 - WinWeb Crawler 2.0 para rastrear la Web del INE.
 - Log Parser 2.2 para el análisis previo de los logs.

- Alter Wind Log es una herramienta utilizada para el análisis de los patrones de navegación de usuarios de un sitio Web.
- Salidas del software Analogs para medidas descriptivas.
- Google Analytics: es un servicio gratuito de estadísticas de sitios Web. Ofrece información agrupada según los intereses de tres tipos distintos de personas involucradas en el funcionamiento de una página: ejecutivos, técnicos de marketing y Webmasters.
- Documentación en Word, Excel, Openproject y diseño de esquemas con iGRAFx.

IDENTIFICACIÓN Y DEFINICIÓN DE SUBSISTEMAS

Los Subsistemas que se identifican en el Sistema "Webclicks-INE" son los siguientes:

- S1.- Exploración de la Web del INE (Fase 0).
- S2.- Obtención de los Ficheros de Salida (Fase1).
- S3.- Desarrollo del aplicativo de preprocesamiento automático (Fase 2).

Se describe la división del Sistema en Subsistemas, de los Subsistemas en Funciones y de estas en Subfunciones. Se describirán también narrativamente cada una de las partes y posteriormente los almacenes de datos (en esta aplicación ficheros planos). Para cada Función o Subfunción identificada en detalle se describirá su relación con las entidades del sistema identificadas, las características propias del proceso (posibles algoritmos, actualización, consultas, etc.), tipo de tratamiento (batch o interactivo) y otro tipo de información como puede ser frecuencia de ejecución, interrelación entre las partes, orden de ejecución....

La descomposición del sistema en subsistemas queda detallada a continuación:

S1.- EXPLORACIÓN DE LA WEB DEL INE (FASE 0).

- **Función 1.1.- "Generación de Ficheros Excel con las URL del *Web Crawler* y el Análisis del *Website* del INE"**
 - **Subfunción 1.1.1.- "Generación Fichero Excel de Rastreo Web".**

Se pasa un rastreador Web (WinWebCrawler 2.0) por toda la Website del INE, obteniendo la taxonomía de las páginas indexables del INE. Se obtiene un fichero excel de 11.446 registros (Entidad D.2.1).

- **Subfunción 1.1.2.- “Generación Fichero Excel de Website del INE”.**

Se rastrea manualmente la Web del INE, relacionando URL a subgrupos identificativos. Dichos grupos se han realizado con un doble criterio, por un lado la propia organización de la Web del INE y por otra la organización funcional para permitir el análisis posterior que se hubiera decidido.

Esta clasificación es necesaria, vista la enorme cantidad de URL indexables (sin mencionar las dinámicas).

Se obtienen 57 grupos identificativos (“Agricultura”, “Aplic_IPC”, “Aplic_Nombres”, “Aplic_Var_IPC”, “ayuda”, “Ayudacod”, “biblioteca”, “buscador”, “calendario”, “CensoAgrario”, “CensoElectoral”, “censos”, “censos2011”, “charts”, “Ciencia_Tecnologia”, “contactar”, “cuentas”, “datos_abiertos”, “datos_internacionales”, “Demografia_Poblacion”, “economia”, “Entorno_Fisico_Medio_Ambiente”, “epa”, “explica”, “fondo_documental”, “FormacionYEmpleo”, “geotempus”, “gescla”, “Grafica_tabulacion”, “home”, “Indicadores_Economicos”, “indice_Web”, “Industria_Energia_Construccion”, “infoeuropea”, “Informacion_INE”, “intercensal”, “ipc”, “masINE”, “Menu_tabulacion”, “meParezco”, “Metodologia_Estandares”, “microdatos”, “Nomenclator”, “prensa”, “Productos_Servicios”, “publicaciones”, “rss”, “SedeElectronica”, “servicios”, “Sintesis_Estadistica”, “sociedad,” “Tab_Menu” (páginas de tabulación de *request* con *referrer* una página de Menu de Tabulación), “Tab_Otros” (páginas de tabulación que no provienen del menú de tabulación), “Tabulacion” (proveniencia desconocida), “tempus”, “territoriales”). Esta es la Entidad D.2.1.

- **Subfunción 1.1.3.- “Generación Fichero Excel de Website del INE a partir del Google Analytics (GA) y el software Analogs”.**

Se busca con el GA las URL más buscadas (Informe-Contenido-Visión General), si bien hay que tener en cuenta que los resultados que ofrecen han tenido que ser contrastados empíricamente para comprobar la coherencia.

Asimismo, se ha accedido a las salidas del Analog y se ha pasado el *software* de Alter Wind Log para comprobar las salidas más frecuentes. Esta es la Entidad D.2.3.

- **Subfunción 1.1.4.- “Generación Fichero Excel de *Request/Referrers* del INE”.**

Tomando como entrada las entidades de las tres subfunciones anteriores, se realiza un cruce entre ambos ficheros y se complementa con las URL no identificadas en la subfunción 1.1.2. Es decir, se

tiene una relación de las URL que nos vamos a encontrar antes de analizar los logs. Se obtiene como salida la entidad D.3.

- **Función 1.2.- “Generación de Ficheros Excel con las IP no válidas”**

- **Subfunción 1.2.1.- “Generación Fichero Texto de IP no válidas”.**

Se obtiene de la Web www.iplists.com la lista de los *Web crawlers* más habituales. El fichero tiene 4.367 registros y corresponde a la Entidad D.4.

Este fichero se contrastará con los logs del INE antes de darlos por válidos.

- **Subfunción 1.2.2.- “Parseamiento de los logs del INE”.**

Con el propósito de buscar los *Web crawlers* de los logs del INE, se usa el *software* LogParser 2.0 para tratar de identificarlos. Asimismo, dicho software sirve para eliminar registros no válidos (por ejemplo, *request* de una hoja de estilo, etc.). Después de un primer análisis, se decide no usar dicho software, para tratar de mantener el control sobre todo parseamiento de la información.

S2.- OBTENCIÓN DE LOS FICHEROS DE SALIDA (FASE1).

El esquema físico de los datos es el representado en el Diagrama 3.3. Desde ahora se desarrolla en SAS y tomando como muestra los logs del 1/4/2013.

- **Función 2.1.- “Generación del Fichero de Secuencias”**

- **Subfunción 2.1.1.- “Lectura y edición primaria de los logs”.**

Tiene como entrada la entidad D1 y como salida la D.6.1. Se leen los datos de los ficheros planos de log y nos quedamos con los válidos. Además esta subfunción añade un identificador al registro. Esto es, nos quedamos con los que tienen Código 2xx o 3xx (2xx: Peticiones correctas: Esta clase de código de estado indica que la petición fue recibida correctamente, entendida y aceptada. 3xx: Redirecciones) y los Métodos GET, POST y HEAD (El método GET requiere la devolución de información al cliente identificada por la URL. El método POST se creó para cubrir funciones como la de enviar un mensaje a grupos de usuarios, dar un bloque de datos como resultado de un formulario a un proceso de datos, añadir nuevos datos a una base de datos, etc. El método HEAD es igual que el método GET, salvo que el servidor no tiene que devolver el contenido, sólo las cabeceras).

Sería un objetivo complementario y no cubierto por este estudio de logs, el análisis de estas peticiones erróneas, aunque se destaca que hay un número significativo.

- **Subfunción 2.1.2.- “Eliminación de Peticiones no válidas”**

Tiene como entrada la entidad D.6.1 y como salida la D.6.2. Se eliminan las peticiones al servidor que no son visitas a páginas que nos interese analizar, como las peticiones de *.jpg, *.gif, etc.

- **Subfunción 2.1.3.- “Eliminación de IP de *Spiders*”**

Tiene como entrada la entidad D.6.2 y como salida principal la D.6.3 limpio de IP de *spiders*, crawlers y otras IP que no tienen un comportamiento que nos interese analizar. Además tiene como entrada auxiliar la D.4 salida del Subsistema anterior y como salida auxiliar el fichero Spiders_ID (D.7.1).

Este fichero debe ser revisado cada vez que corra el programa para comprobar que efectivamente se trata de este tipo de visitas.

Se hace así por motivos de eficiencia de procesamiento, un fichero log de un día puede contener más de cinco millones de registros.

- **Subfunción 2.1.4.- “Eliminación de registros inválidos”.**

Tiene como entrada la entidad D.6.3 y como salida la D.6.4. Para ello tiene como entrada auxiliar la entidad D.5 de URL no válidas. Se ha obtenido de manera recursiva comprobando el error al cargar aquellas URL resultantes de la entidad D.7.2 obtenida en la Subfunción 2.1.5.

- **Subfunción 2.1.5.- “Formación de la variable *Request/Referrer*”.**

Tiene como entrada la entidad D.6.4 y como salida la D.6.5.

Como entrada auxiliar tiene la entidad D.3 de ficheros Request obtenida en el Subsistema 1. Se forma la variable *Request /Referrer* y se obtienen los ficheros de D.7.2 y D.7.3 de Registros en blanco. La salida D.7.2 sirve a su vez como entrada a la Subfunción 2.1.4.

Se blanquean aquellas *request* que hayan quedado en blanco por lo que es necesario la comprobación del fichero Request_Blanco (D.7.2).

Se desagrega el *request* de tabulación en función del peticionario (si viene o no del menú de tabulación).

- **Subfunción 2.1.6.- “Cálculo del *gap* entre hits por visita”**

Tiene como entrada la entidad D.6.5 y como salida la D.6.6.

Se formatea la fecha, se ordena el fichero y se calcula el *gap* entre *hits*, considerando que para una misma IP, son dos visitas distintas si pasa más de media hora de inactividad.

Se calcula una variable de registro por visita y otra variable secuencial por visita y se ordena.

Se obtiene como entidad de salida auxiliar la D.7.4, Fichero de *Spiders* que cumplen que tienen más de 750 hits por visita. Este fichero debe ser comprobado puesto que se eliminan dichas visitas en la siguiente Subfunción.

- **Subfunción 2.1.7.- “Segunda eliminación de *spiders*”.**

Tiene como entrada la entidad D.6.6 y D.7.4 y como salida la D.6.7. Se eliminan todas aquellas *spiders* identificadas en la Subfunción anterior.

- **Subfunción 2.1.8.- “Cálculo del máximo”.**

Tiene como entrada la entidad D.6.7 y como salida la D.6.8.

Se calcula el máximo de *hits* por visita y se ordena el fichero.

- **Subfunción 2.1.9.- “Insertar Fila al inicio”.**

Tiene como entrada la entidad D.6.8 y como salida la D.6.9.

Se inserta una fila correspondiente al primer *hit* por visita y se ordena. Esto se hace para facilitar el análisis, lo mismo que la inserción de la fila “final”.

- **Subfunción 2.1.10.- “Cálculo variable auxiliar sec2”.**

Tiene como entrada la entidad D.6.9 y como salida la D.6.10.

Se calcula una variable auxiliar secuencial para poder renombrar la fila insertada.

- **Subfunción 2.1.11.- “Renombramiento de la fila insertada a Inicio”.**

Tiene como entrada la entidad D.6.10 y como salida la D.6.11.

- **Subfunción 2.1.12.- “Insertar Fila al final”.**

Tiene como entrada la entidad D.6.11 y como salida la D.6.12.

Se inserta una fila correspondiente al último *hit* por visita y se ordena.

- **Subfunción 2.1.13.- “Cálculo variable auxiliar sec3**

Tiene como entrada la entidad D.6.12 y como salida la D.6.13.

- **Subfunción 2.1.14.- “Cálculo variable auxiliar sec4”.**

Tiene como entrada la entidad D.6.13 y como salida la D.6.14.

Se calcula estas variables auxiliares secuenciales para poder renombrar la fila insertada.

- **Subfunción 2.1.15.- “Renombramiento de la fila insertada a Final”.**

Tiene como entrada la entidad D.6.14 y como salida la D.6.15.

- **Subfunción 2.1.16.- “Creación de la variable *ip_def*”.**

Tiene como entrada la entidad D.6.15 y como salida la D.6.16.

Se calcula como la concatenación del IP y el registro de la visita (variable calculada en Subsistema 2.1.6)

- **Subfunción 2.1.17.- “Creación del Fichero de Secuencias”.**

Tiene como entrada la entidad D.6.16 y como salida la D.7.5.

Mantiene sólo las variables de *ip_def*, *sec*, *datetimevar1*, *request* y *referrer*.

- **Función 2.2.- “Generación del Fichero de Page”**

- **Subfunción 2.2.1.- “Estudio de los secs_null”.**

Tiene como entrada la entidad D.7.5 y como salida la D.6.17.

Corresponden a aquellos *referrer* que tienen el valor “NULL” (no tienen valor asignados en los logs). Para este fichero, complementario al “Secuencias” se van a eliminar, pero queda una salida en la librería *work* si se quiere examinar.

- **Subfunción 2.2.2.- “Eliminación de los registros de Referrer=NULL”.**

Tiene como entrada la entidad D.6.17 y D.6.16 y como salida la D.6.18.

- **Subfunción 2.2.3.- “Creación variable auxiliar sec_1”.**

Tiene como entrada la entidad D.6.18 y como salida la D.6.19.

- **Subfunción 2.2.4.- “Correr un puesto los Referrer”.**

Tiene como entrada la entidad D.6.19 y como salida la D.6.20. Debido a la creación de los registros de “inicio” y “final” se hace corresponder a “Inicio” el referrer original y a “final” el último request, corriendo el resto un puesto.

- **Subfunción 2.2.5.- “Creación del Fichero de Page”.**

Tiene como entrada la entidad D.6.20 y como salida la D.7.6.

- **Función 2.3.- “Generación del Fichero de Weblogs41”**

- **Subfunción 2.3.1.- “Obtención de la matriz por registro”.**

Tiene como entrada la entidad D.6.16 y como salida la D.6.21.

Si una visita tiene algún *hit* en alguno de los grupos identificadores se identifica con una nueva variable que valdrá 1 y 0 en caso contrario.

- **Subfunción 2.3.2.- “Creación variable clicks y long”.**

Tiene como entrada la entidad D.6.21 y como salida la D.6.22.

Para cada visita el número de clicks y la duración total de la visita.

- **Subfunción 2.3.3.- “Pasar a numéricas las variables de la Subfunción 2.3.1”.**

Tiene como entrada la entidad D.6.22 y como salida la D.6.23.

- **Subfunción 2.3.4.- “Creación tabla con variables”.**

Tiene como entrada la entidad D.6.23 y como salida la D.6.24. Se obtiene como función MAX de cada variable (así sólo puede valor 1 o 0).

- **Subfunción 2.3.5.- “Creación tabla Weblogs41 con variables sin duplicados”.**

Tiene como entrada la entidad D.6.24 y como salida la D.7.7. Una visita por registro y ordenado por ip_def.

- **Función 2.4.- “Generación del Fichero de Weblogs44”**

- **Subfunción 2.3.1.- “Obtención de la matriz por registro”.**

Tiene como entrada la entidad D.6.23 y como salida la D.6.25.

Si una visita tiene algún *hit* en alguno de los grupos identificadores se identifica con una nueva variable que valdrá 1 y 0 en caso contrario. En este caso se obtiene como función SUM de cada variable (así valdrá para cada variable en número de *hits* por visita).

- **Subfunción 2.3.2.- “Creación tabla Weblogs44 con variables sin duplicados”.**

Tiene como entrada la entidad D.6.25 y como salida la D.7.8. Una visita por registro y ordenado por ip_def.

- **Subfunción 2.3.3.- “Creación tabla cluster con variables sin duplicados”.**

Tiene como entrada la entidad D.7.8 y como salida la D.6.26. Agrupa en Grupos los Subgrupos identificativos de la entidad D.7.7 con una visita por registro y ordenado por ip_def.

- **Subfunción 2.3.3.- “Creación tabla cluster2 con variables sin duplicados”.**

Tiene como entrada la entidad D.6.25 y como salida la D.7.9. Realiza la clasificación definitiva con una visita por registro y ordenado por ip_def.

- **Función 2.5.- “Generación de los Ficheros de Análisis”**

- **Subfunción 2.5.1.- “Obtención de la matriz por registro”.**

Tiene como entrada la entidad D.7.6 y como salidas las D.8.1-D.8.5

Es una muestra de un procedimiento automático de detección de patrones, que no se va a desarrollar más en el presente proyecto. Necesita tener instalados algunos módulos específicos de SAS (el SAS Miner).

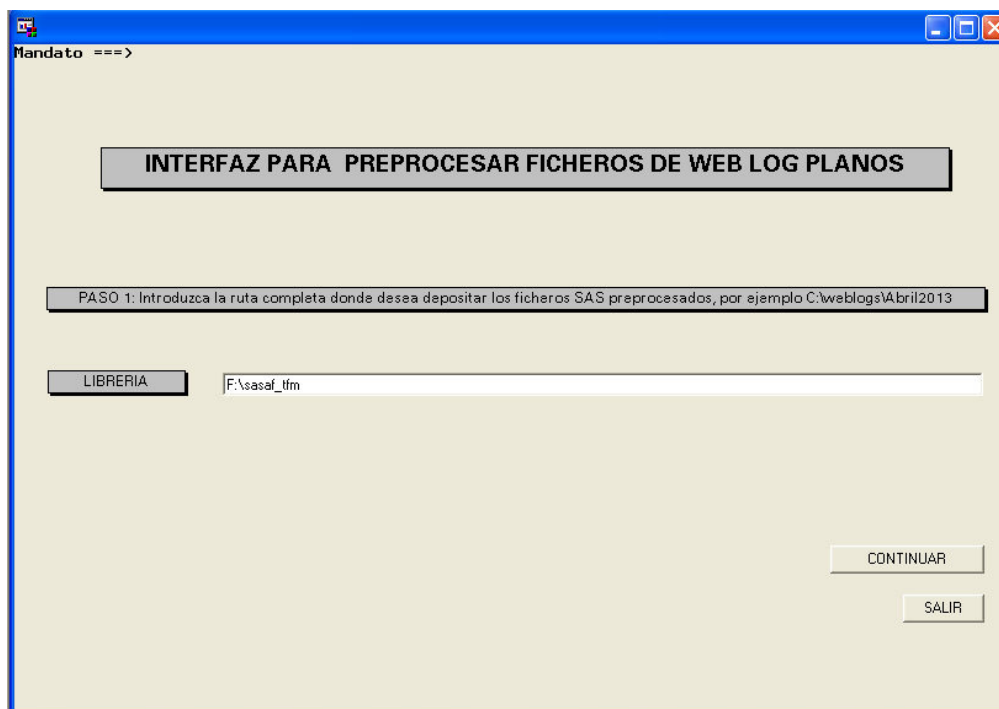
S3.- PROGRAMACIÓN DEL APLICATIVO (FASE 2).

Una vez se dispone de la programación, se realiza una macro y se trabaja sobre el *front-end* del usuario, el aplicativo.

- **Función 3.1.- “Interfaz de usuario”**

- **Subfunción 3.1.1.- “Entrada de librería”.**

El usuario indicará la librería donde quiere depositar los ficheros creados (las entidades D7 y D8), con la posibilidad de continuar y salir de la aplicación.



Imágen 3.1. Impresión de pantalla del primer paso de la interfaz del aplicativo

- **Subfunción 3.1.2.- “Chequeo de ficheros de entrada y procesamiento”.**

La aplicación debe cumplir que se puedan procesar todos los *logs* que se desee. En la práctica, en esta arquitectura de PC y dada la cantidad de recursos necesarios, se ha limitado el proceso de ficheros a 7, que corresponder a una semana si cada log contiene los datos de un día. Es importante indicar que los *logs* del INE, en este momento contienen más de cinco millones de *hits* diarios.

Por otro lado, por eficiencia de procesamiento esta aplicación trata los ficheros por separado y luego los une, luego es necesario que cada fichero contenga los datos del *Weblog* al menos de un día entero. Técnicamente, sería posible procesar un fichero con todo el contenido del *Weblog* mensual por ejemplo. Ahora bien, se insiste en que su procesamiento llevaría varias horas y que no sería necesario en relación al propósito de análisis.

En este apartado el usuario deberá indicar los ficheros que quiere procesar, es decir, las entidades D1. Desde un fichero hasta siete.

Se realizará un control de errores para que el usuario cumplimente en orden y marcando la ruta completa, incluido el tipo de fichero, en nuestro caso “.txt”

Después se pasa el proceso. Para ello se ha creado una macro a partir de la programación del subsistema 2.

Imágen 3.2. Impresión de pantalla del segundo paso de la interfaz del aplicativo

○ **Subfunción 3.1.3.- “Control de Ficheros”.**

El usuario tendrá la posibilidad de navegar por las librerías y ficheros creados, además de poder filtrar por la variable de interés, a través de menús o directamente con comandos de *query*.

	ip	request	req_s	sec		
1	10.58.79.170	inicio	1	0	10	
2	10.58.79.170	rss	1	1	10	
3	10.58.79.170	rss	1	2	10	
4	10.58.79.170	rss	1	3	10	
5	10.58.79.170	final	1	4	10	
6	10.58.80.117	inicio	2	0	10	
7	10.58.80.117	SedeElectronica	2	1	10	
8	10.58.80.117	territoriales	2	2	10	
9	10.58.80.117	home	2	3	10	
10	10.58.80.117	Tab_Otros	2	4	10	
11	10.58.80.117	prensa	2	5	10	
12	10.58.80.117	home	2	6	10	
13	10.58.80.117	home	2	7	10	
14	10.58.80.117	CensoElectoral	2	8	10	
15	10.58.80.117	Metodologia_Estandares	2	9	10	
16	10.58.80.117	Tab_Otros	2	10	10	
17	10.58.80.117	SedeElectronica	2	11	10	
18	10.58.80.117	territoriales	2	12	10	
19	10.58.80.117	home	2	13	10	
20	10.58.80.117	home	2	14	10	
21	10.58.80.117	Tab_Otros	2	15	10	

Imágen 3.3. Impresión de pantalla del tercer paso de la interfaz del aplicativo

ARQUITECTURA DE LA INFORMACION

Para el Sistema “WEBClicks INE”, hemos determinado la **Matriz asociación de Entidades/Funciones** (Cuadro 3.1) y la **Matriz asociación de Funciones/Áreas** (Cuadro 3.2).

CUADRO 3.1 MATRIZ FUNCIONES/ENTIDADES (E=ENTRADA, S=SALIDA)

ENTIDADES	D.1 Ficheros de Logs	D.2 URL del Web Crawler/Análisis del Web	D.3 Ficheros de Request y Referrer	D.4 Ficheros de IP no válidas	D.5 Ficheros de URL NO Validas	D.6 Resto Ficheros Work SAS Weblogs	D.7 Ficheros SAS Page, Weblogs41, Spider, Request_Blanco, Cluster2	D.8 Ficheros SAS Análisis Web
S.1 -FASE 0	E/S	E	S	E				
S.2 -FASE 1	E		E/S	E/S	E/S	S	S	
S.3 - FASE 2	E		E/S	E/S	E/S	S	S	S

CUADRO 3.2.- MATRIZ FUNCIONES/ÁREAS INVOLUCRADAS

ÁREAS FUNCIONES	DESARROLLO	DIFUSION	SISTEMAS
S.1 -FASE 0	X	X	
S.2 -FASE 1	X		X
S.3 - FASE 2	X	X	

ESQUEMA FÍSICO DE LOS DATOS.

Las entidades pueden consultarse en el apartado anterior de “Arquitectura de la información”. En relación al esquema físico de los datos del Sistema queda reflejado en el siguiente Figura 3.1, recogiendo su relación así como la obtención de cada Almacén de Datos en el tiempo.

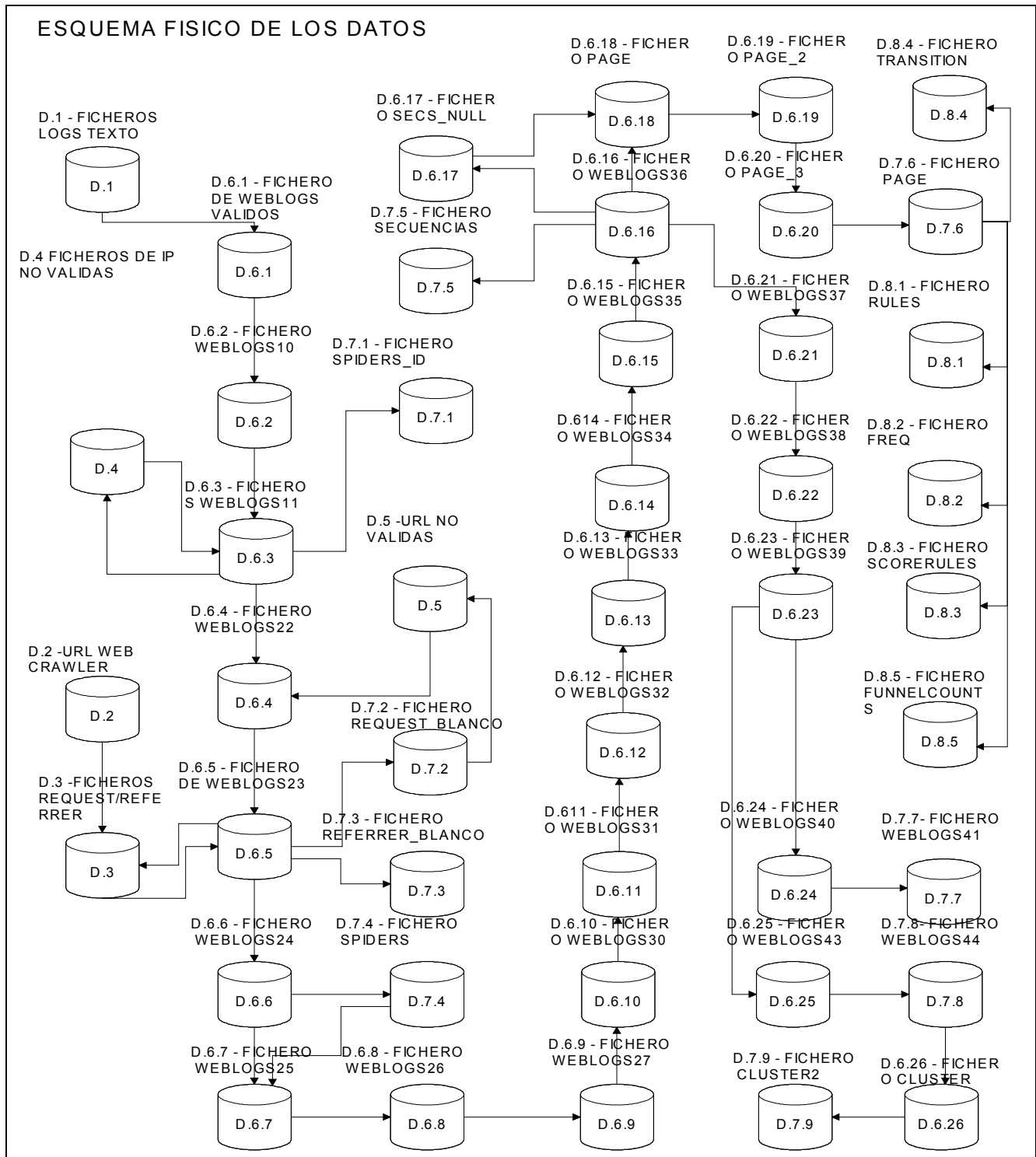


Figura 3.1. Esquema físico de los datos del Sistema de Información

DESCRIPCIÓN DE LOS DATOS

Se ha empleado un diseño de investigación exploratorio, justificado por su novedad y originalidad y dada la ausencia de antecedentes de investigaciones de esta naturaleza dentro de la Administración Pública.

Los datos seleccionados son los que se extraen de los servidores Web del INE entre el 24 de Abril y el 30 de Abril, para capturar el efecto ante la publicación de las estadísticas con más repercusión del INE, como son la EPA (publicado el 25 de abril), el IPC (29 de abril) o el PIB avance (30 abril). Las variables de los ficheros de log son los que se explican en el anexo en el apartado de LOGS.

El número de observaciones de cada fichero es el de la siguiente tabla, donde:

FECHAS	PAGE	SECUENCIAS	WEBLOGS41	WEBLOGS44
24/04/2013 (1)	212.171	576.004	49.956	49.956
25/04/2013 (2)	247.549	752.621	62.519	62.519
26/04/2013 (3)	188.533	510.267	44.853	44.853
27/04/2013 (4)	109.267	315.225	28.084	28.084
28/04/2013 (5)	139.673	287.572	30.760	30.760
29/04/2013 (6)	227.103	531.969	50.999	50.999
30/04/2013 (7)	201.947	569.264	47.669	47.669
TOTAL	1.326.243	3.542.922	314.840	314.840

Tabla 3.1. Totales de los Ficheros según el día y total.

El fichero Page incluye los Peticionarios (*Referrer*) de las peticiones (*Request*) y si estaba en blanco se ha eliminado toda la visita. Por ello, el fichero resulta ser aproximadamente el 40% del que procede, el fichero de Secuencias.

Los ficheros Weblogs (41 y 44), conteniendo el mismo número de registros, nos indican el número de visitas, según el criterio utilizado (una misma visita debe ser del mismo IP y con menos de media hora de inactividad). Así el 25 de abril hubo 62.519 visitas al *Website* del INE.

Si ahora consideramos las visitas por mismo IP en esos días (visitantes únicos), sin considerar el tiempo de inactividad, tendríamos los siguientes registros de la tabla 3.2:

FECHAS	24/04/2013 (1)	25/04/2013 (2)	26/04/2013 (3)	27/04/2013 (4)	28/04/2013 (5)	29/04/2013 (6)	30/04/2013 (7)	TOTAL
MISMO_IP	32.423	43.134	30.338	18.554	20.791	34.169	31.682	167.440

Tabla 3.2. Visitas con mismo IP

De esta manera resultan 43.519 visitantes únicos para un mismo IP en ese mismo día. Si lo comparamos con los datos del Google Analytics (GA) observamos una discrepancia de alrededor del 15% para los visitantes únicos (las visitas son calculadas con otra metodología). Se destaca la diferencia en la metodología (el concepto de “visita” cambia) y en la posibilidad de que el cliente restrinja las cookies, con lo que se reduce consecuentemente el número de visitas del GA.

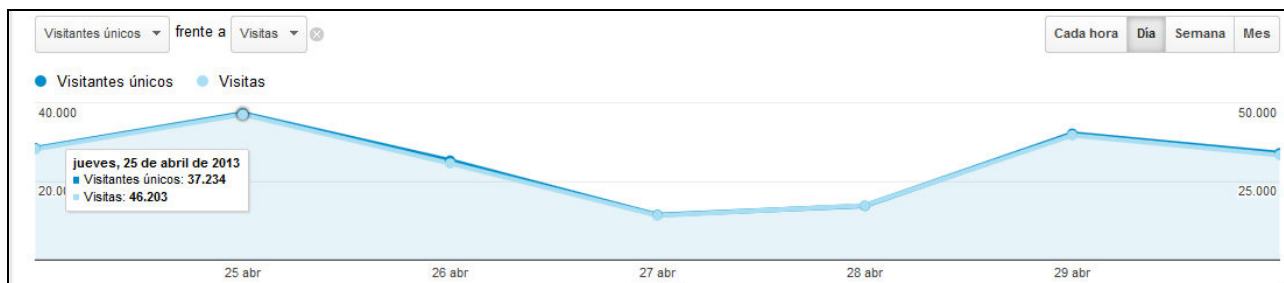


Gráfico 3.1 Visitas y Visitas Únicas según Google Analytics.

El **FICHERO SECUENCIAS** contiene, según las especificaciones anteriores, las variables de *ip_def* (ip modificada para ser clave única), *request* (la petición) y *sec* (la secuencia dentro de la visita). Un detalle del mismo se puede apreciar en la tabla 3.3:

	ip_def	request	sec
1452613	195.53.62.25.11759	inicio	0
1452614	195.53.62.25.11759	Ayudacod	1
1452615	195.53.62.25.11759	Ayudacod	2
1452616	195.53.62.25.11759	Ayudacod	3
1452617	195.53.62.25.11759	Ayudacod	4
1452618	195.53.62.25.11759	Ayudacod	5
1452619	195.53.62.25.11759	final	6
1452620	195.53.62.25.11760	inicio	0
1452621	195.53.62.25.11760	Ayudacod	1
1452622	195.53.62.25.11760	Ayudacod	2
1452623	195.53.62.25.11760	Ayudacod	3
1452624	195.53.62.25.11760	Ayudacod	4
1452625	195.53.62.25.11760	Ayudacod	5
1452626	195.53.62.25.11760	Ayudacod	6
1452627	195.53.62.25.11760	final	7

Tabla 3.3. Fichero transaccional de Secuencias

Se han introducido para análisis los registros de Inicio (al principio de la visita) y Final (al final de la visita).

Se trata de un fichero transaccional en que el orden (la secuencia) es importante y que se ha derivado a partir del campo de fecha. Este fichero contiene todos los *hits* para este periodo de tiempo y se usará para determinar reglas de asociación y secuencia.

Alternativamente se ha elaborado un *dataset* denominado **PAGE**, utilizado para ir haciendo el seguimiento de la ruta seguida por el visitante, un detalle del cual se puede ver a continuación:

	ip_def	sec	request	referrer_final
39789	190.27.182.209.12146	0	inicio	externo
39790	190.27.182.209.12146	1	Menu_tabulacion	inicio
39791	190.27.182.209.12146	2	Menu_tabulacion	Menu_tabulacion
39792	190.27.182.209.12146	3	Menu_tabulacion	externo
39793	190.27.182.209.12146	4	Menu_tabulacion	externo
39794	190.27.182.209.12146	5	final	Menu_tabulacion
39795	190.27.182.7.12147	0	inicio	externo
39796	190.27.182.7.12147	1	Aplic_Nombres	inicio
39797	190.27.182.7.12147	2	final	Aplic_Nombres

Tabla 3.4. Fichero transaccional de Page

En cada registro del log, hay un campo que es el Referrer_final (campo de *referrer* modificado) y otro que es el Request, que indican para cada petición de dónde viene (Referrer_final) y qué solicita (request).

Otra vía para hacer esto hubiera sido ir relacionado *requests* subsiguientes sin considerar el campo de *referrer*, pero de esa manera se perdía la información del *referrer*. Por ejemplo un visitante puede refrescar la página o puede tener varias sesiones abiertas. Esto se puede ver en el detalle de la tabla 3.4, donde el *referrer* es externo al Website del INE en varias ocasiones en una misma visita.

El listado detallado de request se puede consultar en el apartado de Análisis Funcional en la Identificación y Definición de Subsistemas, Subfunción 1.1.2. Hay muchas particularidades respecto al Request que no se detallan en este trabajo, baste con indicar que a un servidor se le pueden hacer *Requests* de páginas que no existen, por ejemplo: ine.es/noexiste, que quedan guardados en el servidor. Por eso ha sido necesaria la Fase 0, donde se han identificado las URL válidas y es una posible explicación a las diferencias con el GA.

Para la elaboración del `Referrer_final` se ha tenido en cuenta, como particularidad respecto a la elaboración del `request`, que el campo de `Peticionario` viene con la identificación completa de la URL (el `www`) y que los que venían en blanco se han transformado a `NULL` (aunque en `PAGE` no afecte ya que según lo comentado se eliminan del fichero). Además, se han considerado nuevos Subgrupos identificativos, como el `Intranet` (que no puede ser `request` pero sí `referrer`) y `Externo` (si el `referrer` es de un dominio externo al INE).

El fichero de [WEBLOGS41](#) ya comentado, contiene para cada `ip_def` (uno por visita, aunque se mantiene el de visitante único para análisis), el número de clicks, la hora de inicio y la duración de la visita (si es un click la duración es desconocida y se deja a 0, lo que arrastra a la duración global a la baja) y para cada subgrupo identificativo si ha sido o no visitada (1 si lo ha sido y 0 si no).

	ip	ip_def	reg_s	clicks	inicio	duracion	home	buscador	fondo_documental	Informacion_INE
1	1.179.150.237	1.179.150.237.1	1	10	1682409647	743	0	0	0	1
2	1.187.75.25	1.187.75.25.2	2	1	1682436268	0	0	0	0	0
3	10.58.10.56	10.58.10.56.3	3	19	1682414236	508	1	0	0	0
4	10.58.10.7	10.58.10.7.4	4	1	1682412416	0	0	0	0	0
5	10.58.10.7	10.58.10.7.5	5	176	1682418329	9253	0	0	0	0
6	10.58.10.7	10.58.10.7.6	6	19	1682430130	362	0	0	0	0
7	10.58.10.7	10.58.10.7.7	7	38	1682432597	2876	0	0	0	0
8	10.58.108.104	10.58.108.104.8	8	1	1682429261	0	0	0	0	1
9	10.58.108.138	10.58.108.138.9	9	2	1682429564	1072	1	0	0	0
10	10.58.108.195	10.58.108.195.10	10	4	1682439330	104	1	1	0	0
11	10.58.11.119	10.58.11.119.11	11	1	1682417977	0	1	0	0	0
12	10.58.11.160	10.58.11.160.12	12	1	1682411932	0	1	0	0	0
13	10.58.11.24	10.58.11.24.13	13	11	1682428240	139	0	0	0	0

Tabla 3.5. Fichero de Weblogs41

En este *dataset*, en análisis adicionales se podrían incluir otras variables binarias relacionadas con las visitas del usuario.

Por último, el fichero de [WEBLOGS44](#), contiene para cada `ip_def` (el de visita, aunque se mantiene el de visitante único para análisis), el número de clicks, la hora de inicio y la duración de la visita (si es un click la duración es desconocida) y para cada subgrupo identificativo el número de veces que ha sido visitada.

	ip	ip_def	reg_s	clicks	inicio	duracion	home	buscador	fondo_documental
1	1.179.150.237	1.179.150.237.1	1	10	1682409647	743	0	0	0
2	1.187.75.25	1.187.75.25.2	2	1	1682436268	0	0	0	0
3	10.58.10.56	10.58.10.56.3	3	19	1682414236	508	1	0	0
4	10.58.10.7	10.58.10.7.4	4	1	1682412416	0	0	0	0
5	10.58.10.7	10.58.10.7.5	5	176	1682418329	9253	0	0	0
6	10.58.10.7	10.58.10.7.6	6	19	1682430130	362	0	0	0
7	10.58.10.7	10.58.10.7.7	7	38	1682432597	2876	0	0	0
8	10.58.108.104	10.58.108.104.8	8	1	1682429261	0	0	0	0
9	10.58.108.138	10.58.108.138.9	9	2	1682429564	1072	1	0	0
10	10.58.108.195	10.58.108.195.10	10	4	1682439330	104	1	2	0

Tabla 3.6. Fichero de Weblogs44

IV.- ANALISIS Y EJECUCIÓN DEL PROCEDIMIENTO DE MINERÍA WEB

Este Trabajo considera cómo el comportamiento de los usuarios de un *Website* puede ser predicho desde el análisis de los datos existentes considerando el orden en que se han visitado las páginas.

Cuando un usuario hace un Link en un *Website* el servidor se queda la traza de sus acciones en un *logfile*, y esto se denomina *clickstream*. Cada *hit* del usuario, en términos globales, se corresponde con una petición de una página web. Una sesión de usuario (o visita) describe la sucesión de páginas que son vistas durante un tiempo logeado a un *Website* determinado.

El objetivo del análisis es mostrar cómo el *clickstream* puede ser usado para extraer las rutas de navegación más usadas en un *Website* para poder predecir (incluso *online*) qué otras páginas va a visitar el usuario, dada la ruta seguida hasta ese momento. Esto se puede usar también para buscar la probabilidad de que un usuario entre en una determinada página o para ver la probabilidad de que se salga del *Website*, entre muchas posibilidades.

A continuación, se va a realizar el proceso de minería de datos: Análisis Exploratorio de los datos, Construcción del Modelo, Comparación de los Modelos e Informe de Conclusiones.

ANÁLISIS EXPLORATORIO DE DATOS

El objetivo es descubrir las reglas de secuencias más comunes entre las 57 variables binarias del Fichero Weblogs41, describiendo si un subgrupo identificativo de páginas ha sido visitado. Para obtener conclusiones válidas, el *dataset* debe ser homogéneo y para determinarlo se hace un análisis exploratorio de los datos.

Se va a analizar las distribuciones univariantes de los clicks, duración e hora de inicio.

Variable	Rol	Media	Desviación estándar	No ausente	Ausente	Mínimo	Mediana	Máximo	Asimetría	Curtosis
clicks	INPUT	9.10683	26.06156	100000	0	1	3	500	10.28129	140.4921
duracion	INPUT	614.1909	1979.466	100000	0	0	19	39611	7.884593	86.67416
hora_inicio	INPUT	135172.9	59179.11	100000	0	0	135659	235959	-0.40783	-0.47836

Tabla 4.1. Medidas descriptivas de las variables “clicks”, “duracion” y “hora_inicio”.

Para **CLICKS** se observa que son 9,25 clicks por visita (frente a 7,84 del GA), siendo el valor mediano de 2 y el valor máximo de 749. Esto es porque por convención y para discernir sólo los

comportamientos de interés para el estudio, se han considerado como *spiders* y eliminado si eran más de 750 hits por visita. Este sería un valor parametrizable en la programación.

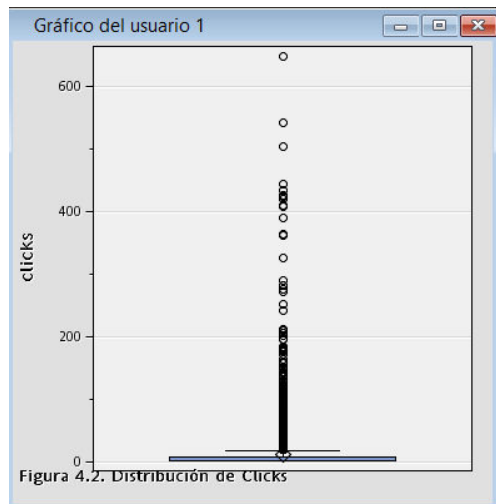


Gráfico 4.1. Distribución univariante de la variable “clicks”

Para **DURACION** se observa que son 632 segundos por visita (es decir, 10 minutos y medio frente a casi 6 del GA), siendo el valor mediano de 15 segundos y el valor máximo de 86.439 (24 horas). Este sería un valor parametrizable en la programación.

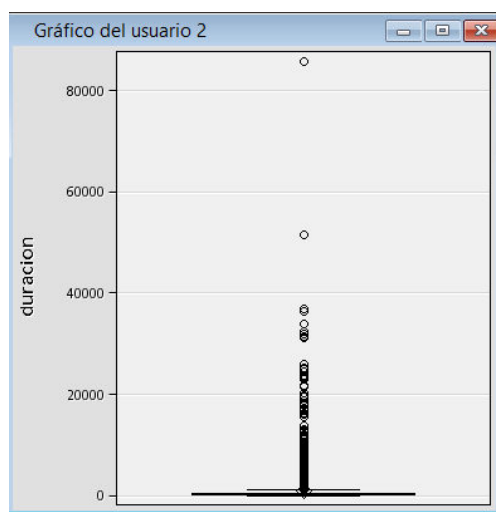


Gráfico 4.2. Distribución univariante de la variable “duracion”

Para **HORA_INICIO** se observa que la distribución tiene dos picos, con una distribución similar a la del GA y un pico de entradas hacia las 13:50, excepto en una entrada de peticiones muy alta en pocos minutos desde las 00:00.

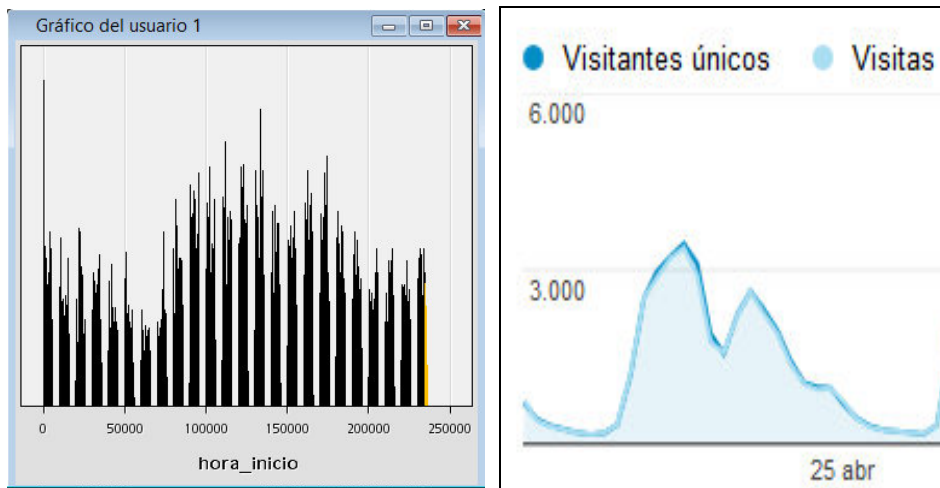


Gráfico 4.3. Figura del 24 de abril de densidad de la variable “hora_inicio” (en formato numérico, por ejemplo las 957 corresponden a las 00:09:57) comparada con la del GA (segundo gráfico).

La variable de duración sí que nos refleja un comportamiento anómalo por exceso de algunos registros, a pesar de su tratamiento específico durante la programación. Asimismo, la variable “hora_inicio” también refleja un inicio masivo a finales del día, lo que tampoco sería razonable.

Hay que mencionar que para su tratamiento ha sido necesario convertir la hora en numérico, por lo que para los minutos y segundos se hace necesaria la conversión de vuelta desde numérico.

Se procede a eliminar los *outliers* para la variable “click” y “duracion”. En concreto, si el número de clicks es superior a 500 y si la duración es superior a 40.000 segundos (11 horas). Corresponde a 635 visitas, que son eliminadas de todos los ficheros, quedando lo siguiente:

FECHAS	PAGE	SECUENCIAS	WEBLOGS41	WEBLOGS44
TOTAL	1.318.070	2.977.176	314.208	314.208

Tabla 4.2. Totales de los Ficheros sin *outliers*.

Ahora los estadísticos cambian a lo siguiente:

Variable	Rol	Media	Desviación estándar	No ausente	Ausente	Mínimo	Mediana	Máximo	Asimetría	Curtosis
clicks	INPUT	8.347095	22.48406	314208	0	1	2	500	10.02563	143.2446
duracion	INPUT	541.6605	1823.117	314208	0	0	15	39952	9.082764	116.3485
hora_inicio	INPUT	135127	60353.56	314208	0	0	135257	235959	-0.4132	-0.51951

Tabla 4.3. Medidas descriptivas de las variables “clicks”, “duracion” y “hora_inicio”, con el fichero sin *outliers*.

Se hace a continuación un análisis descriptivo de las 57 variables binarias en el dataset depurado:

Variable	%	Variable	%	Variable	%
Home	30,874	censos2011	2,188	Entorno_Fisico_Medio_Ambiente	0,608
Menu_tabulacion	27,736	explica	1,844	microdatos	0,603
Metodologia_Estandares	24,958	Nomenclator	1,730	Agricultura	0,576
Charts	21,080	economia	1,635	intercensal	0,574
RSS	12,995	epa	1,485	Síntesis_Estadística	0,560
Tab_Menu	11,278	servicios	1,479	Ciencia_Tecnología	0,524
Prensa	10,157	censos	1,446	Tempus	0,457
Productos_Servicios	9,626	FormacionYEmpleo	1,212	datos_internacionales	0,443
Informacion_INE	7,433	cuentas	1,201	Infoine	0,425
Grafica_tabulacion	5,864	biblioteca	1,123	MasINE	0,389
Buscador	5,752	Industria_Energia_Const	1,062	contactar	0,296
Aplic_Var_IPC	5,661	fondo_documental	1,054	CensoAgrario	0,235
Tab_Otros	5,631	ayuda	1,019	infoeuropea	0,234
Demografia_Población	5,471	territoriales	0,997	calendario	0,229
Aplic_IPC	5,461	SedeElectronica	0,955	indice_Web	0,210
Tabulacion	5,393	ipc	0,892	Gescla	0,076
Publicaciones	4,404	Ayudacod	0,00835	geotempus	0,00050
Sociedad	4,183	CensoElectoral	0,00810	meParezco	0,00015
Aplic_Nombres	3,786	Indicadores_Economicos	0,00788	datos_abiertos	0,00002

Tabla 4.4. Análisis descriptivo de las 57 variables

Destaca “home” con un 31% de frecuencia (“Charts” es la petición de chart que se hace automáticamente cuando se carga la página Home llegando hasta el 21%), “Menu_Tabulacion” con un 27% y “Metodologia_Estandares” con un 25%. Después hay una bajada pronunciada hasta el 13% de “rss”, 11% de “Tab_Menu” y 10% de “prensa”. Esto nos dice que de todas las visitas, en el 31% se pasa por home, por ejemplo (es su Suporte).

Podríamos suponer que existe un comportamiento diferencial entre los usuarios que entrar a la Web a través de la página de inicio del INE, “home” y el resto, que entran a cualquier página del *Website* del INE distinta a home directamente.

Home	observaciones	variable	Media
0 (no visita home)	217.199	clicks	5,45
		duracion	458,97
		hora_inicio	13:15:58
1 (visita home)	97.009	clicks	14,83
		duracion	726,81
		hora_inicio	13:51:27
Total	314.208	clicks	8,35
		duracion	541,67
		hora_inicio	13:31:05

Tabla 4.5. Medias condicionales de las variables cuantitativas con respecto a “home”.

Se puede observar que la hora de entrada al *Website* es aproximadamente la misma, pero cuando se entra por "home" se realizan casi tres veces más de hits y se permanece un 20% más en el Website. Una explicación lógica sería que el usuario si entra a través del "home" debe hacer más hits hasta llegar a la página que desea visitar.

Dada la heterogeneidad de los visitantes se va a hacer un análisis *cluster*, concretamente el procedimiento K-medias apropiado para muestras de elevado tamaño, con el fin de encontrar conglomerados con comportamientos similares. Las variables *cluster* que se consideran son “hora_inicio”, “clicks” y “duracion” por el método de Ward, con máximo preliminar de 50 y mínimo de 40 *clusters*:

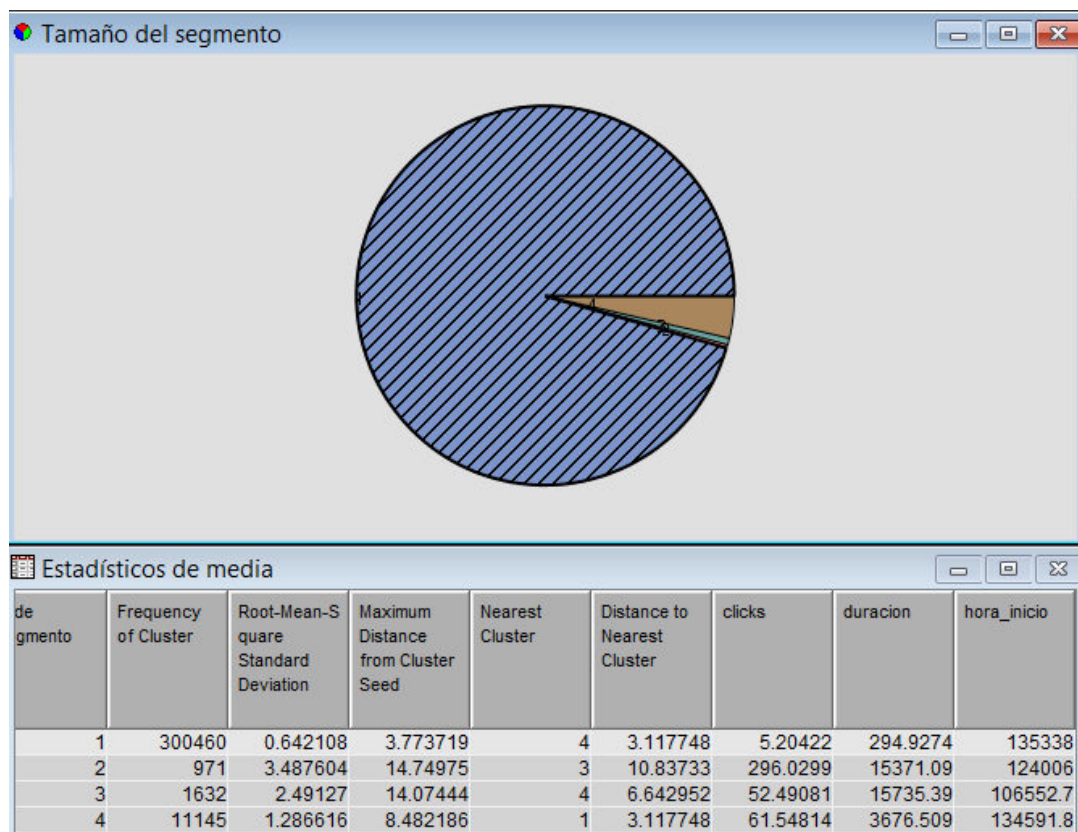


Gráfico 4.4. Clusters sobre el total de la muestra, sin outliers

A partir de un análisis de resultados, se decide usar el cluster principal en lo sucesivo (Id=1), que contiene el 95% de las observaciones pero concentradas entorno a los 5 clicks y los 5 minutos (menos clicks y menos tiempo que comparado con el GA) y que, además, explica el 95% de la R^2 (de la salida de resultados no mostrado en este informe).

Cluster	obs	Variable	Media
1	300460	Clicks	5,2
		Duración	294
		hora_inicio	13:35:27
Total	314208	clicks	8,35
		duracion	541,67
		hora_inicio	13:31:05

Tabla 4.6. Medias condicionales de las variables cuantitativas con respecto al cluster.

Realizando de nuevo la segmentación se obtiene resultados interesantes, en especial considerando que es la hora de inicio la variable con más importancia:

Importancia de la variable				
NAME	LABEL	NRULES	NSURROGATES	IMPORTANCE
hora_inicio		4	8	1.00000
duracion		8	7	0.94656
clicks		9	3	0.86596

Tabla 4.7. Importancia de las variables en la “clusterización” del primer conglomerado

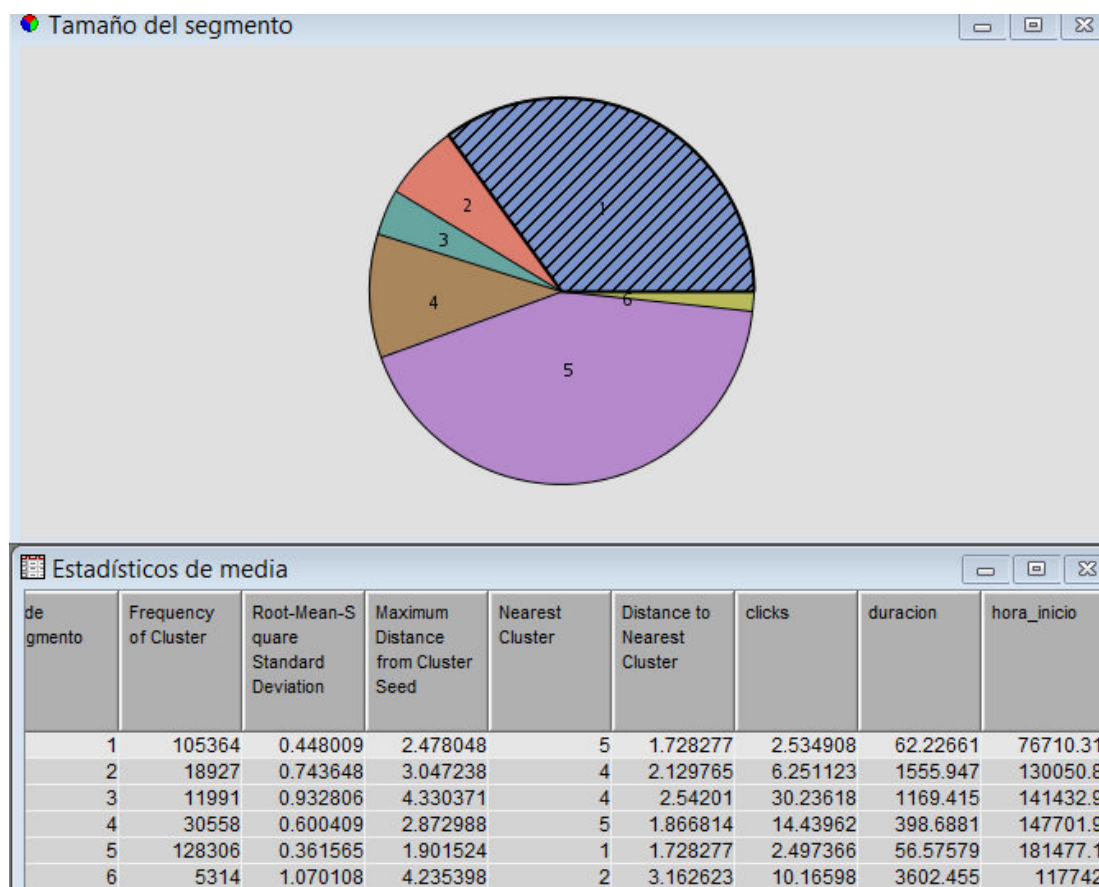


Gráfico 4.5 Clusters sobre la muestra principal (Cluster 1 anterior) y sin outliers

Lo que podríamos llamar "a primera hora de la mañana" hay una cantidad muy importante de visitas (aproximadamente el 30%) con pocos clicks y corta duración (un minuto) y el cluster más importante es aquel con hora de inicio las 18 horas igualmente con pocos clicks y poca duración (un minuto).

Las visitas más largas son aquellas que se producen por la mañana, especialmente justo antes de mediodía, y las que tienen mayor cantidad de hits son las iniciadas entorno a las 14 horas.

El fichero Weblogs44 se usaría complementariamente para agrupar los visitantes a partir de las páginas vistas y la frecuencia con que se visitan esas páginas sin distinguir por duración, hora de inicio o número de clicks totales. De esta manera, a partir de la media normalizada podemos discernir por comportamiento, por ejemplo grupos que sólo visitan un área determinada del *Website* (más especializados), otros que visitan un grupo de áreas determinadas, otros que sólo irán a la parte de ayuda, etc. Para ello, se hace necesario agrupar todos los Subgrupos (los 57) en otros de mayor tamaño, que identifican:

- Si el usuario busca información del INE, de sus productos, del empleo que ofrece (Grupo O1infoine)
- Si busca datos de INEBASE o microdatos (Grupo O2inebase)
- Si es perfil de prensa o rss (Grupo O5prensa)
- Si busca información metodológica (Grupo O3metodologiaestandares)
- Si hace uso de la ayuda, búsquedas o contacta a "infoine" (Grupo ayuda) o
- Si está interesado por las aplicaciones que ofrece el INE (Grupo aplicaciones).

Aplicando un análisis Cluster con el método de Ward, que minimiza la varianza intragrupo y maximiza la intergrupos, con corte *Cubic Clustering Criterion* (CCC) de 3, se llega al grupo mayoritario 3, que podemos caracterizar como “usuario veleta”, poco especializado que hace pocos clicks en cualquiera de los grupos (menos de uno en media).

Por su parte, el grupo 5, “Usuario de Prensa”, está especializado en Prensa. El 1, “Usuario Profesional” está interesado en Metodología junto a un uso intensivo de inebase y de información del INE. Es similar a los grupos 2 y en especial el grupo 4, “Usuarios que no encuentra o que busca más información”, que además necesitan ayuda.

Por último, el grupo 6, el “visitante lúdico” básicamente "juega" con las aplicaciones del INE con un uso intensivo, lo cual nos hace pensar que son aplicaciones interesantes para el usuario pero que no está interesado en el resto de información estadística más “tradicional”.

Esta información podría sugerir apoyar cierto tipo de información durante determinadas franjas horarias.

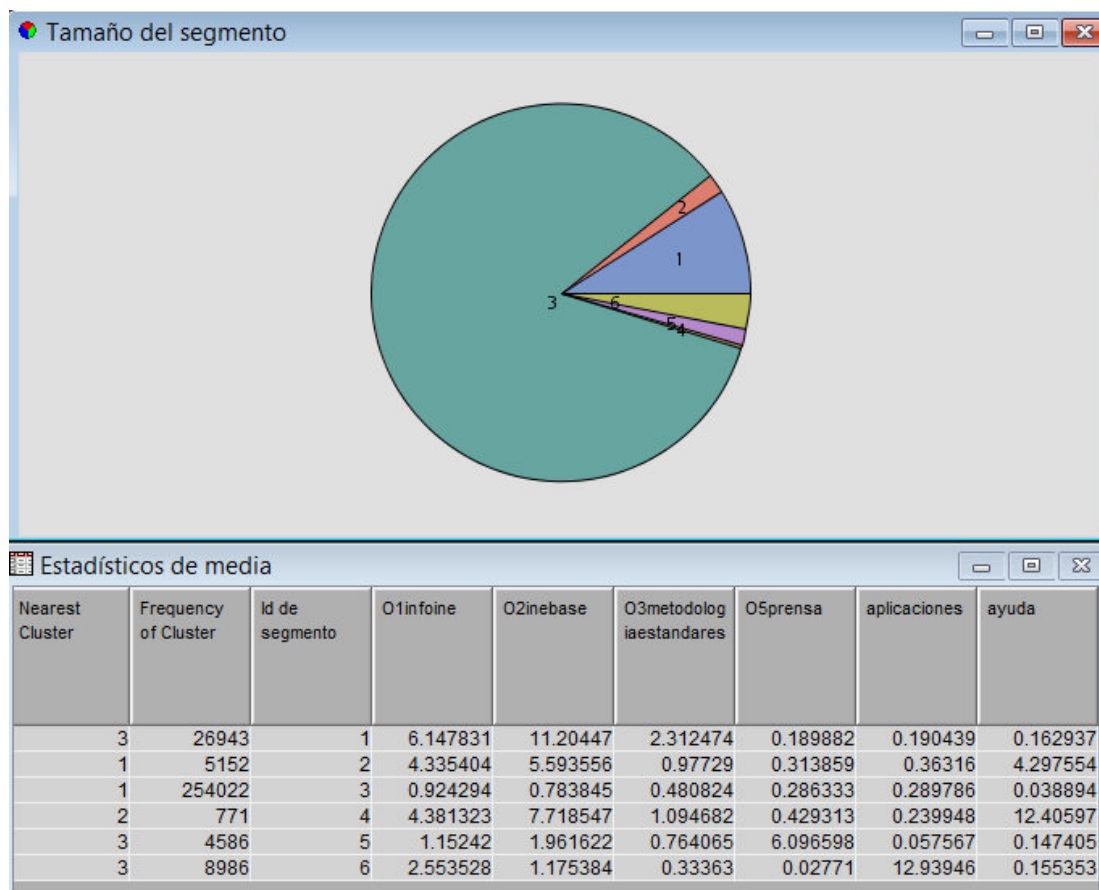


Gráfico 4.6. Clusters sobre la muestra principal (*Cluster 1* anterior) y sin *outliers* a partir del fichero Weblogs44.

Es evidente que el usuario "prototipo" de la Web no hace un uso intensivo del mismo, en su mayoría uno o dos clicks por sesión. Por otro lado, puede ser muy interesante centrar los esfuerzos en futuros estudios en el grupo que necesita pedir ayuda para encontrar lo que necesita, con el objetivo de investigar qué necesita el usuario y por qué no lo ha encontrado en su búsqueda (problema de estructura, de contenido o de conocimiento del usuario). Como una aproximación, vamos a realizar una exploración de las búsquedas que realizan los usuarios a través del buscador, que queda reflejado en el log en la petición (*request*) realizada.

Para ello, nos creamos el fichero de Búsquedas con las búsquedas realizadas en un día determinado (24 de abril, día de publicación del IPRI) resultando unas 11.000 búsquedas. Se hace una exploración a partir del nodo de *Text Mining*, desechando preposiciones y dando estructura al fichero para el análisis.

La relación lineal de la siguiente figura indica que es proporcional las veces que salen los textos objeto de análisis por documento (búsqueda).

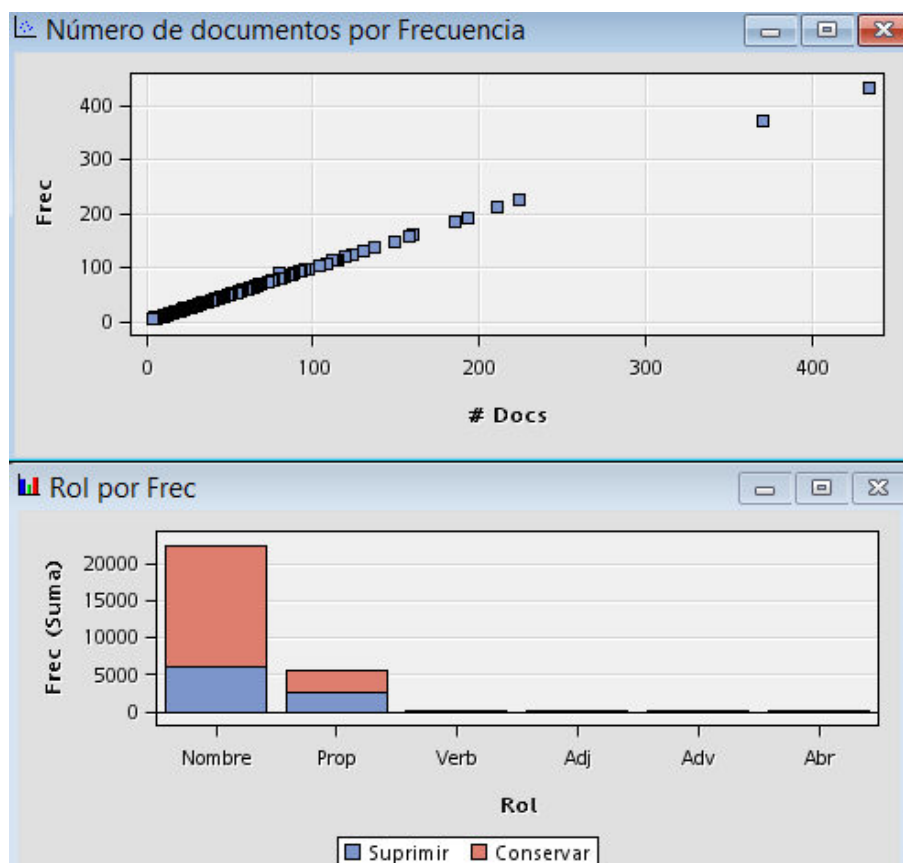


Gráfico 4.7. Frecuencia de documentos y rol diferenciando si se conservan

En la figura anterior se aprecia que algunos textos se destacan frente al resto. De las más de 11.000 búsquedas, más de cuatrocientas corresponden a "población" y "tasa de ipc" y en orden decreciente las siguientes: "paro", "poblaci", "pib", "ipc", "pib", "consumo", "renta", "edad", "actividad", "empleo", "mortalidad", "densidad", "sector", "persona", "index", "vivienda", "epa", "gasto", "delitos", "encuesta", "sectores", "trabajo", "extranjeros", "parados", "capita", "car", "sexo", "precios", "madrid", "numero", "alquileres", "natalidad", "indice", "municipios", "cnae", "violencia", "epa" (frecuencia de 76).

Este análisis adquiriría mayor interés si se amplía el análisis a un mayor rango temporal (más días analizados) y relacionando la búsqueda con el momento de la visita en que lo busca (ítem dentro de su ruta). Es decir, porqué busca el término y en qué momento de la visita. En esta primera aproximación, la conclusión puede ser que las búsquedas son sobre los productos "estrella" del INE, es decir Población, IPC, EPA y Cuentas nacionales, pero hay que destacar que con menos frecuencia hay términos en las búsquedas que indican que se asocia al INE como un "proveedor" universal de información, apareciendo así términos de búsqueda como Canarias, euribor, precio del agua, diabetes, universidad...

REGLAS DE SECUENCIA

Para alcanzar los objetivos especificados anteriormente y para el tipo de datos de los que se dispone, se empieza por aplicar un modelo de asociación local.

Dado que las transacciones que se consideran aquí son sesiones de visita, el orden en que se visitan las páginas es importante. Por lo que consideraremos la aplicación de Reglas de Secuencia, que son reglas de asociación ordenadas por una variable, que en nuestro caso es la variable "Sec", que indica la secuencia temporal obtenida durante la programación. Además, se han introducido, según lo especificado anteriormente los registros para cada Visita con *request* "Inicio" y "Final", ajustando el *referrer* según el caso.

Como se ha indicado, se va a considerar sólo el *cluster* de interés y sin *outliers*. Se ha aplicado el algoritmo de SAS de "Asociación", estableciendo un porcentaje de soporte para la asociación de 0,05, que especifica la frecuencia de transacción mínima para permitir una asociación.

Este algoritmo usa las siguientes medidas para evaluar las reglas de asociación:

- Soporte - El nivel de Soporte indica la frecuencia con que la asociación ocurre en la base de datos transaccional. En otras palabras, el Soporte cuantifica la probabilidad de que una transacción contenga a ambos elementos A y B.

Matemáticamente, la regla de asociación $A \Rightarrow B$ se puede expresar como el ratio entre las transacciones que contiene a A y B y todas las transacciones.

- Confianza - La fuerza de una asociación se define por su coeficiente de Confianza. Dada la asociación $A \Rightarrow B$, la Confianza es la probabilidad condicionada de que la transacción contenga a B, dado que la transacción ya contiene al elemento A.

Matemáticamente, sería el ratio entre las transacciones que contienen a A y B y las transacciones que contienen a A.

- Confianza esperada (CE) - Dada la regla de asociación $A \Rightarrow B$, la CE es la proporción de todas las transacciones que contienen a B. La diferencia con la anterior es que éste es una medida del cambio en el poder predictivo causado por la presencia de A en la transacción.

La CE proporciona la indicación de lo que la Confianza debería ser si no hubiera relación entre A y B, es decir es el ratio entre las transacciones que contienen a B y todas las transacciones.

- Lift - Dada la asociación $A \Rightarrow B$, el Lift de la regla de asociación se define como el ratio de la Confianza de la regla entre la CE de la regla. Es decir, es el factor por el que la Confianza excede la CE. A mayor Lift, las asociaciones serán más interesantes (mayor será la influencia de A en la probabilidad de que B esté contenida en la transacción). Lift se puede usar como una medida general de afinidad entre dos elementos de interés (A y B).

Una regla de asociación que nos interese tendrá una Confianza y un Soporte elevado y un nivel de Lift mayor que uno.

Como resultado del algoritmo, se puede obtener las reglas de secuencia con mayor interés empezando por las secuencias indirectas de orden 2 sobre el fichero de Secuencias:

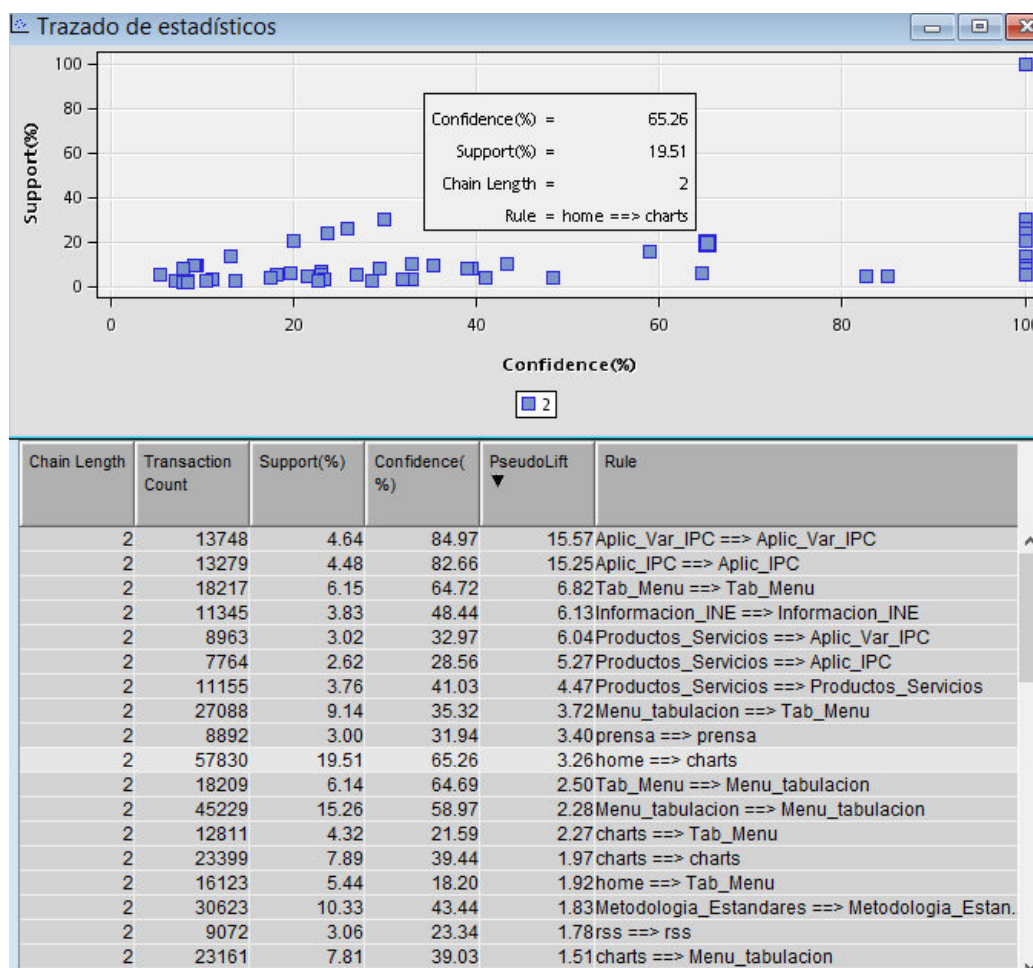


Gráfico 4.8. Secuencias indirectas de orden 2 sobre el fichero Secuencias ordenada por su Lift.

Observamos que las reglas con mayor Lift son las de repetir la petición en las aplicaciones del IPC. Además, con una gran Confianza, debido en parte a que la transacción se prolonga más de un click y a que se hacen peticiones continuadas, seguramente más por interés que por dificultad de utilización.

La regla "Tab_menu"==>"Tab_Menu" nos indica que cuando un usuario entra a ver una tabla, más adelante en la misma sesión ese usuario consulta más tablas. Algo similar sucede con la regla de "Menu_Tabulacion"==> "Menu_Tabulacion", si bien para este caso es usual en el *Website* que se requiera de dos pasos hasta llegar a las tablas deseadas.

La regla "Menu_Tabulacion"==> "Tab_Menu" nos indica que con gran Confianza y como es natural, cuando se entra a un menú de tabulación se entra a continuación a sus tablas correspondientes. Esta es la primera regla, ordenada por Lift, con un nivel alto de Soporte, es decir una regla de especial interés en cuanto a su repercusión.

La regla "home"==>"charts" es consecuencia de que automáticamente cuando el visitante entra a la página de "home", se carga el chart correspondiente, si bien esta relación no se cumple siempre (su Confianza no es el 100%) sino en dos terceras partes de los casos por motivos varios, por ejemplo, que la configuración del dispositivo cliente no permita su carga.

Se puede ir razonando con la misma lógica para el resto de reglas.

Ahora se puede hacer lo mismo, buscar las reglas de asociación indirectas pero de cualquier orden, obteniendo el conjunto de reglas siguiente, que confirma lo que se acaba de exponer:

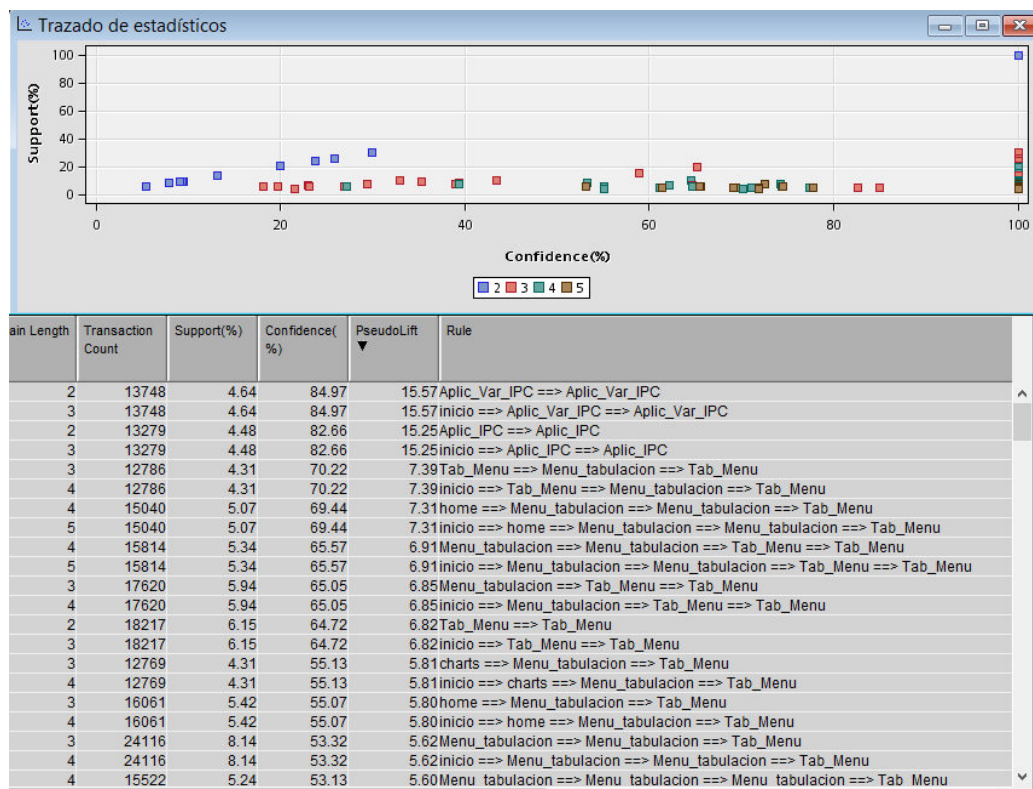


Gráfico 4.9. Secuencias indirectas de cualquier orden sobre el fichero Secuencias ordenada por su Lift.

Se destaca la regla de “home”==>”Menu_tabulacion”==>”Tab_Menu” como una regla presumiblemente lógica dentro del *Website* y que de hecho tiene un alto nivel de Lift, Soporte y Confianza.

A continuación se examinan las secuencias directas de orden 2:

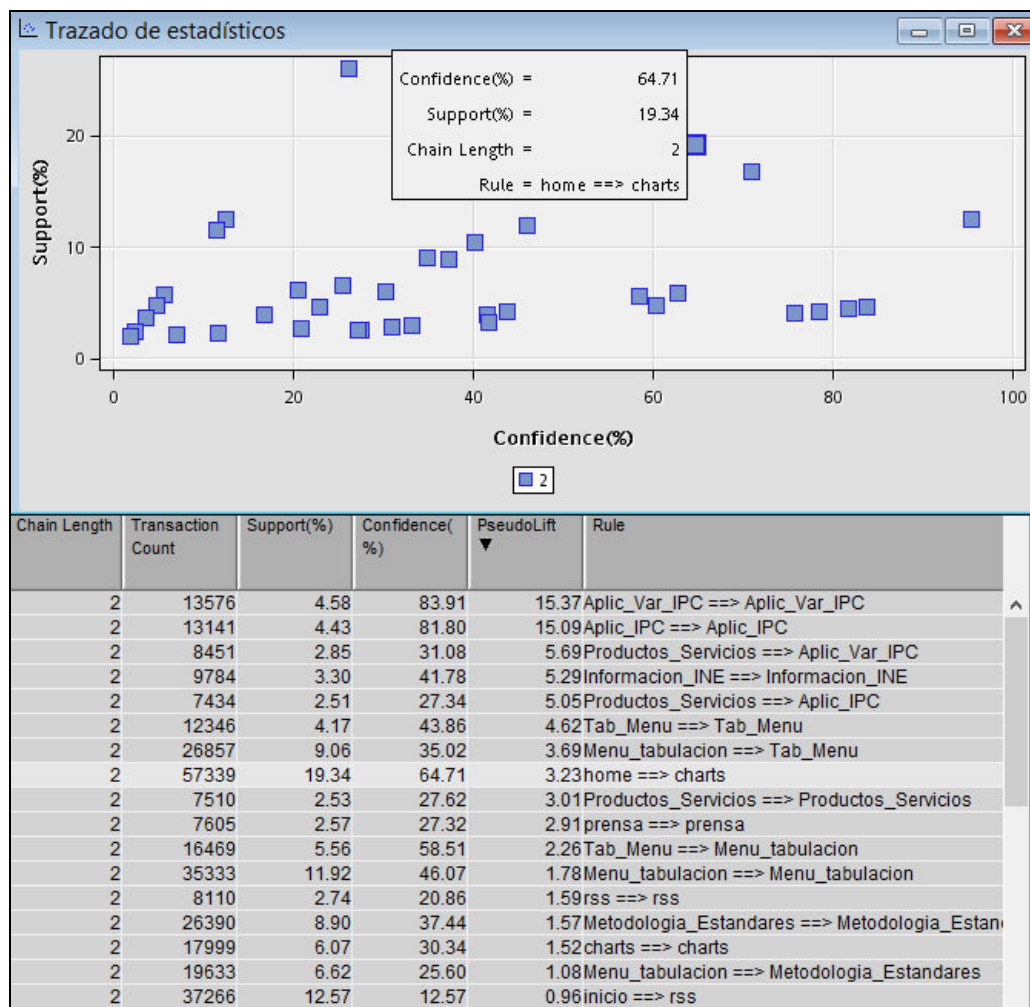


Gráfico 4.10. Secuencias directas de orden 2 sobre el fichero Secuencias ordenada por su Lift.

Hay ciertas variaciones respecto a lo que encontrábamos con las reglas indirectas, por ejemplo desaparece la relación de “home” (www.ine.es)=>”Tab_menu” (tabulación desde el menú tabulación), que no es posible de manera directa.

Con la asociación directa podemos establecer cuales son las páginas de inicio y final más habituales de entrada a la Web del INE. Por ejemplo, en la siguiente tabla podemos ver que la Confianza de que después de algún “rss”, como es lógico dada la naturaleza de la transacción, se termine la visita es del casi 100%. A nivel de análisis nos interesa más la relación entre las “aplicaciones_IPC” y “Final”. También destaca que después de una visita a “Prensa”, casi dos terceras partes de los usuarios abandonan el *site* del INE. Las relacionadas con la tabulación se explican más con que el usuario haya encontrado el objeto de su visita.

Transacciones	Soporte (%)	Confianza (%)	Regla
37.142	13	96	rss ==> final
12.624	4	79	Aplic_IPC ==> final
12.280	4	76	Aplic_Var_IPC ==> final
50.039	17	71	Metodologia_Estandares ==> final
17.495	6	63	prensa ==> final
14.150	5	60	Informacion_INE ==> final
11.710	4	42	Tab_Menu ==> final
30.885	10	40	Menu_tabulacion ==> final
9.021	3	33	Productos_Servicios ==> final
13.599	5	23	charts ==> final
18.196	6	21	home ==> final

Tabla 4.8. Secuencias directas de orden 2 de fin de sesión sobre el fichero Secuencias ordenada por su Confianza.

De manera similar, centrándonos en las páginas de inicio, en la siguiente tabla podemos ver que la Confianza de que “rss” sea la página de inicio de sesión, dado que hay un “rss”, es del casi 100%. “Home” es inicio en el 88% de las ocasiones en que hay “home” en la visita, como parece natural y destaca que a las aplicaciones de IPC se entre muy a menudo desde un dominio externo al INE (alrededor del 40%), similar a las páginas de “Metodología_Estándares” y de “Información_INE”, donde se entra casi en la mitad de las veces desde el exterior. También destacan las entradas directas a “prensa”, lo cual no sorprende porque suelen ser entradas programadas.

Transacciones	Soporte (%)	Confianza(%)	Regla
37.266	13	96	inicio ==> rss
77.540	26	88	inicio ==> home
17.152	56	62	inicio ==> prensa
14.245	5	52	inicio ==> Productos_Servicios
34.357	12	49	inicio ==> Metodologia_Estandares
10.711	4	46	inicio ==> Informacion_INE
34.300	12	45	inicio ==> Menu_tabulacion
7.141	2	44	inicio ==> Aplic_IPC
5.950	2	37	inicio ==> Aplic_Var_IPC

Tabla 4.9. Secuencias directas de orden 2 de inicio de sesión sobre el fichero Secuencias ordenada por su Confianza.

ANALISIS DE LINKS

Ahora consideramos cómo tomar las reglas de secuencias y usar un análisis de *links* para buscar un modelo global. En el SAS Miner, el *Link Analysis* tiene como *input* las secuencias establecidas en las reglas de asociación elegidas.

Vamos a considerar dos análisis, uno con las reglas directas y otro con las indirectas, de cualquier orden hasta 5 como máximo, manteniendo como en el apartado anterior el umbral de Soporte al 0,05.

El *Link Analysis* considera cada una de las secuencias como una fila y cuenta el número de filas con una secuencia determinada. A continuación, se puede ver un gráfico con este análisis donde cada página se representa por un nodo y un *link* se dibuja entre los nodos si la cuenta de la secuencia correspondiente de orden 2 no es nula (es decir, si existe esa secuencia). El gráfico nos explica los nodos que están conectados y los que no, siendo la anchura del *link* lo que nos indica la magnitud de la cuenta, es decir la Confianza:

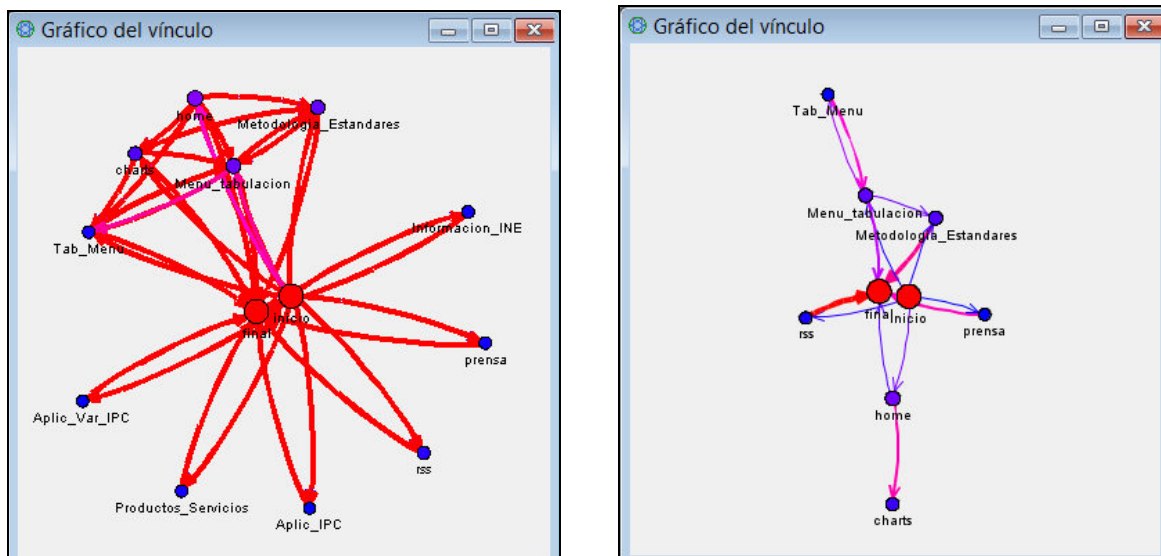


Gráfico 4.11. Gráficas con el *Link Analysis* para secuencias indirectas (izquierda) y directas (derecha) de orden 5 como máximo a partir del fichero de Secuencias.

Por ejemplo, considerando el gráfico de secuencias directas, destaca el *link* entre “home” y “charts”, entre “rss” y “final”, entre “Menu_Tabulación” y “Tab_Menu” y así sucesivamente, lo que significa que son secuencias que aparecen con frecuencia en el *dataset*.

Los *links* se puede observar que son dirigidos. Para orientar la dirección, se toma el conteo mayor entre $A \Rightarrow B$ y $B \Rightarrow A$. Cuando hay una paridad sustancial aparecen dos direcciones, como en el caso de “Menu_Tabulación” y “Tab_Menu”.

La medida de los nodos depende de las medidas de centralidad, que es una idea proveniente de las redes sociales. Una medida de centralidad de primer orden (C1) significa que la importancia del nodo depende del número de conexiones que tenga. Por otro lado, una medida de centralidad de segundo orden (C2) significa que la importancia del nodo depende del número de conexiones que el nodo conectado a él tenga. Utilizamos una medida de centralidad ponderada de primer orden.

La posición del nodo también depende del conteo, ya que el dicho conteo se pone en una matriz de proximidad entre cada par de páginas. Después se usa escalamiento multidimensional para reproducir estas proximidades a través de la distancia Euclídea, dando lugar a dos puntos (coordenadas). A mayor conteo, más cerca en el gráfico cartesiano.

Considerando el fichero de Page, donde según lo explicado en apartados anteriores, se considera para cada petición (*request*) la página de la que proviene dicha petición (*referrer*), tendríamos lo siguiente:

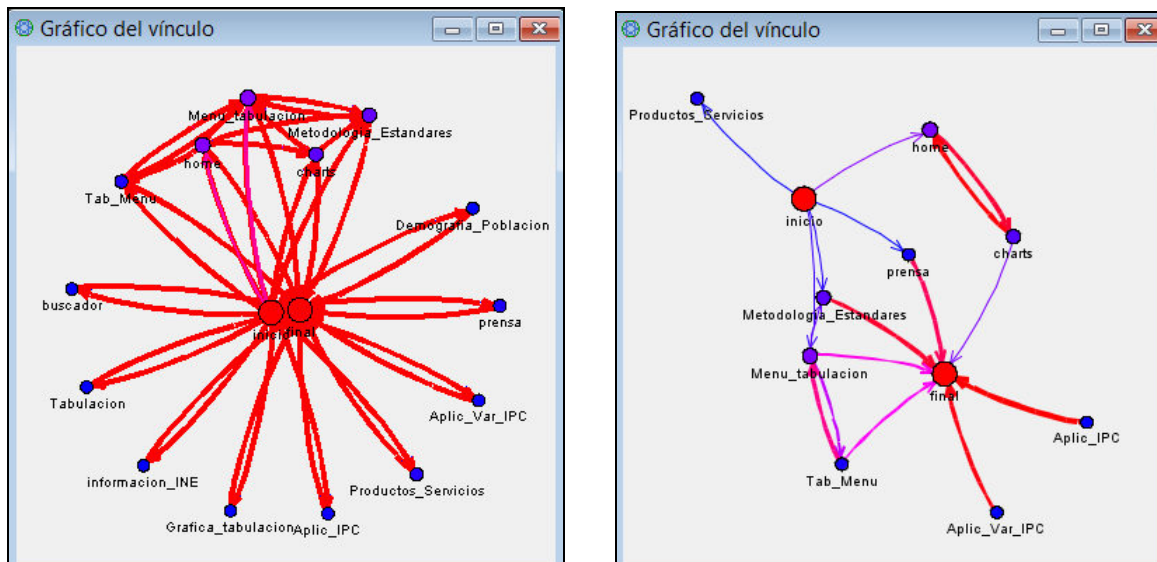


Gráfico 4.12. Gráficas con el *Link Analysis* para secuencias indirectas (izquierda) y directas de orden 5 como máximo a partir del fichero de Page.

Aparecen nuevos *links*, de trazados más anchos y en nuevas posiciones. Hay que considerar que este modelo, que veremos más adelante con más detalle, se trata de un modelo global ya que correspondería a un modelo de cadena de Markov de orden 4, donde se tiene en cuenta el conjunto de secuencias y no sólo las que cumplen cierto requisito. Otra diferencia es que considera el fichero Page, donde se realiza un procesamiento informático adicional al de Secuencias.

SISTEMAS EXPERTOS PROBABILÍSTICOS (SEP)

Estos Sistemas construyen un modelo global a partir de sucesivas factorizaciones. Aunque es muy similar a las reglas de secuencia, la diferencia es sustancial.

Según lo dicho hasta ahora, las reglas de secuencia determinan que la visita a la página B depende de la visita a la página A si el Soporte de la regla $A \Rightarrow B$ es mayor que el umbral prefijado (en nuestro caso 0,05).

En los SEP la variable aleatoria binaria B depende de los sucesos de la variable aleatoria binaria A si $P(B|A, \text{Otras}) \neq P(B|\text{Otras})$, siendo "Otras" el resto de variables consideradas.

Por tanto los SEP son modelos globales por la dependencia entre variables, mientras que las reglas de asociación son modelos locales por la dependencia entre la ocurrencia de eventos. En nuestro contexto, los valores de *request* "inicio" y "final" no son variables aleatorias y no aparecerán en el modelo, que contendrá al resto de valores de *request*. Otra diferencia sustancial es que los SEP discretos se calculan usando tablas de contingencia por lo que no es fácil considerar su temporalidad.

Para comparar los SEP y las reglas de asociación, se ajustará un modelo SEP al *dataset* Secuencias. Esta metodología no está implementada en SAS Miner, así que se va a construir usando una sucesión de regresiones logísticas. El problema a esta aproximación es que no existe un ordenamiento previo de variables, excepto la que consideramos variable objetivo.

El modelo gráfico en la siguiente figura se obtiene considerando las páginas con mayor Soporte, en concreto aquellas a partir de un 2%. Además se prescinde de "charts" y "rss" dada su especial naturaleza y se incluye por motivos analíticos los subgrupos de "epa", "Cuentas" e "IPC". Se considera a la página relevante como de tipo Objetivo y el resto como variables explicativas (de Entrada) a partir del fichero de Weblogs41 "clusterizado":

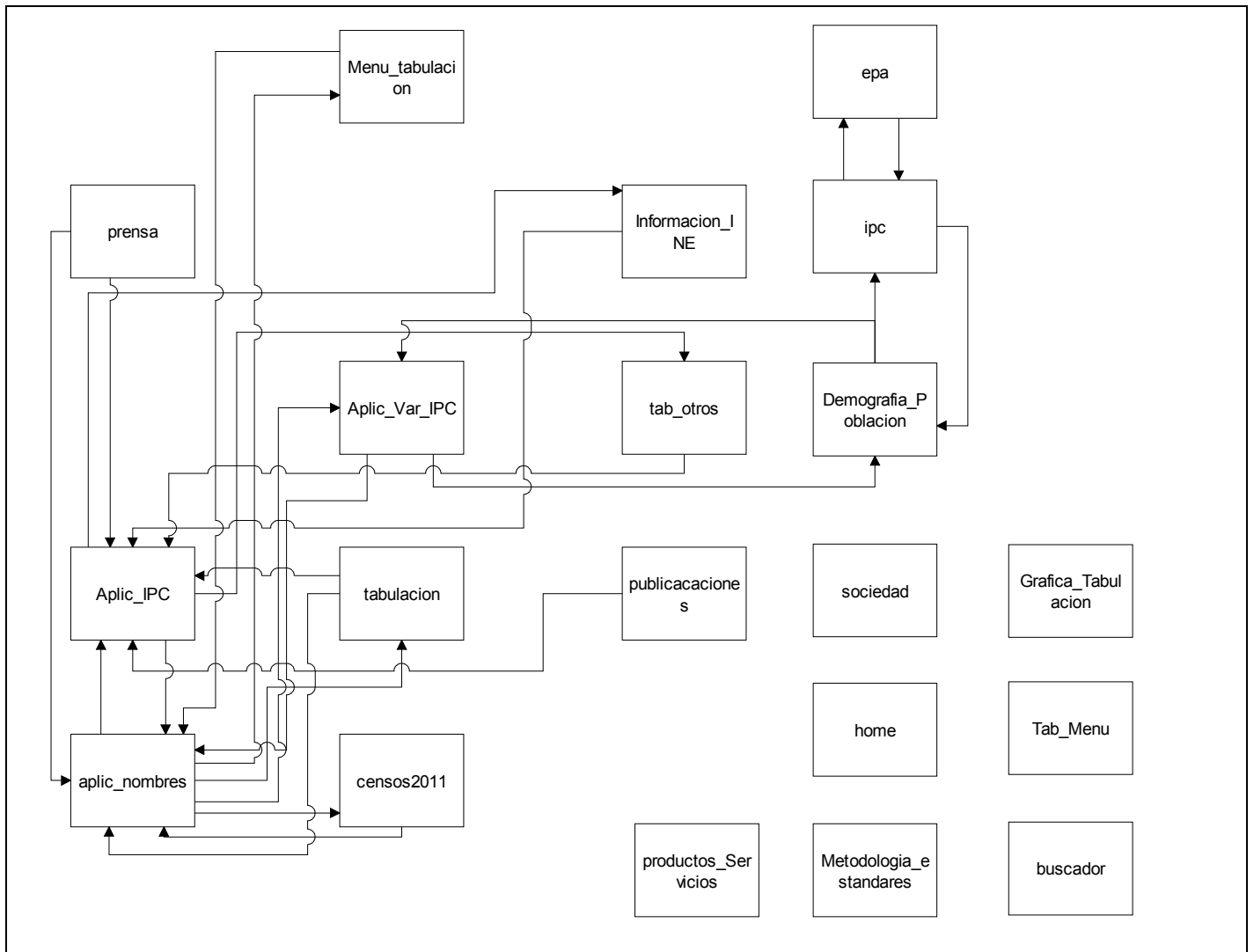


Figura 4.1. Grafo dirigido construido a partir de regresiones logísticas sucesivas.

Destacan con un *odds Ratio* por encima de 10, las relaciones: “epa”==>”IPC”; “Censos2011”==>”Aplic_Nombres”; “Aplic_Nombres”==>”Censos2011”; “Aplic_Nombres”==>”Tabulacion”.

El subgrupo “Buscador” no tiene *odds ratio* por encima de 5. Tampoco lo tiene “Grafica_Tabulacion”, “home”, “Metodologia_Estandares”, “Productos_Servicios”, “Sociedad”, “Cuentas” y “Tab_Menu”.

La Figura 4.1 se puede comparar con los resultados obtenidos en las reglas de secuencias anteriores. Como se ha indicado, los SEP no pueden contener "Inicio" y "Final" por no ser variables aleatorias, por lo que esas secuencias no aparecen en estos grafos. Además los SEP no pueden contener reglas del tipo $A \Rightarrow A$ al estar basados en un *dataset* sin orden, lo que se pone más de manifiesto comparándolas con las reglas indirectas.

Los SEP no consideran el orden de la visita por lo que indica las páginas que influyen a la página objetivo, pero pueden ser antecesoras o sucesoras de la misma. Hay que indicar que para que fuera un SEP puro sería necesario conocer el orden y esto conllevaría que no existirían dobles flechas (de entrada y salida).

El modelo se construye a partir de los resultados de las regresiones logísticas, suponiendo que **si hay un *odds ratio*** (es una medida de la fuerza de asociación entre dos variables binarias) **positivo significativo** (en este estudio se ha establecido un valor de 5), **consecuentemente habrá un *link* entre la variable explicatoria más relevante a la variable objetivo**. El *link* se representa en el gráfico con una flecha desde la variable explicatoria a la variable objetivo. Dado que no hay un ordenamiento *a priori* de las variables, no se ha obtenido un modelo único sino un grafo compatible con distintos ordenamientos.

Por ejemplo, para la variable “Informacion_INE” se obtiene el siguiente trazado de efectos con sus correspondientes estimadores de *Odds Ratio*. Como se puede comprobar el primer efecto positivo corresponde con el mayor *odds ratio*, en este caso “Aplic_IPC”.

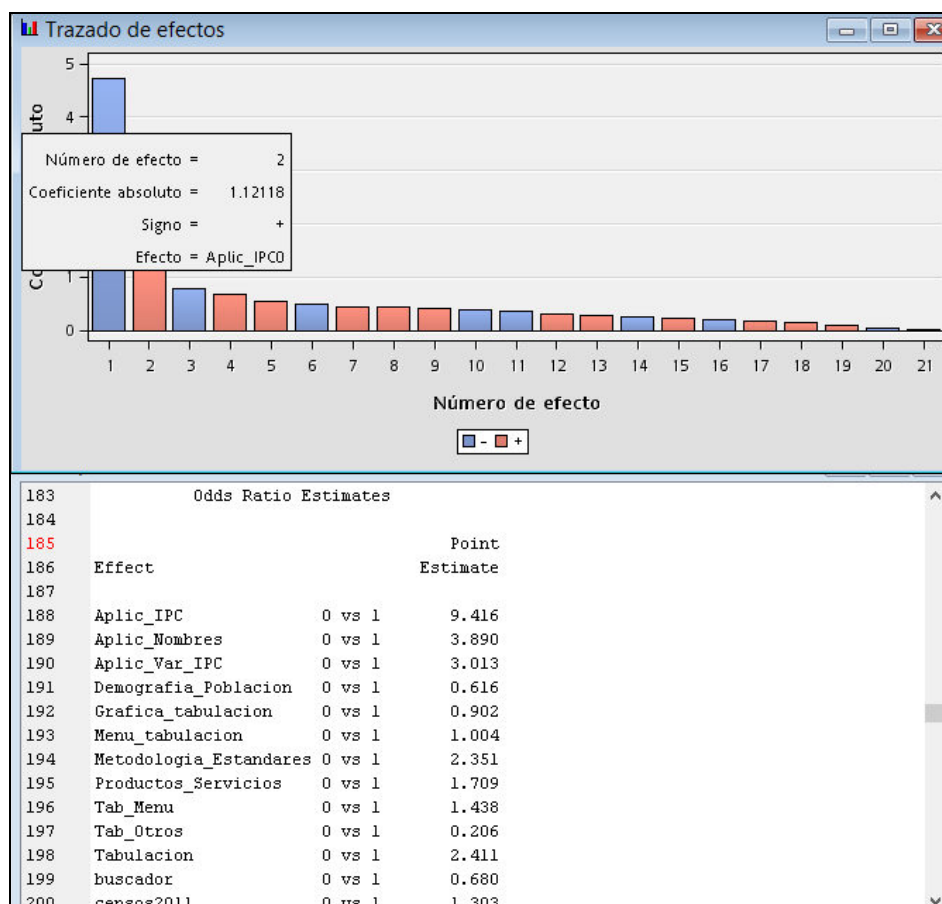


Gráfico 4.13. Trazado de efectos y estimadores *Odds Ratio* para la variable “Informacion_INE”

CADENAS DE MARKOV

El anterior es un modelo global para analizar la dependencia entre variables; son distintos de las reglas de secuencia, que modelizan la dependencia entre eventos. Los modelos de cadena de Markov, en este caso discretas, se pueden usar como modelos globales.

La idea es introducir la correlación entre variables que tengan dependencia del tiempo (la secuencia). En cada sesión, para cada punto del tiempo i , es decir, para el i -ésimo click, se corresponde una variable aleatoria discreta con tantos niveles como número de páginas, que se denominan estados de la cadena. La página i -ésima es la i -ésima realización de la cadena en el momento i , para esa sesión. El tiempo puede ir desde $i=1$ hasta $i=T$, siendo T finito y una sesión puede acabar antes de T , en cuyo caso la última página vista sería un estado absorbente ("Final" en nuestro caso).

Una cadena de Markov establece la dependencia estadística entre lo que se ve antes del momento i (el *referrer*) y lo que se ve en i (el *request*). En particular, una cadena de Markov de primer orden, que es el modelo considerado aquí, establece que lo que se ve en i depende sólo de lo que se ve en $i-1$.

Esta dependencia de corto plazo se puede medir con una matriz de transición que establece la probabilidad de ir de cualquier página a otra en un sólo paso. Para 57 páginas corresponderían 57×57 probabilidades.

Las probabilidades condicionadas de la matriz de transición se pueden estimar a partir de las frecuencias condicionadas.

Asumiendo que la matriz de transición es constante en el tiempo (homogeneidad en la cadena de Markov) se pueden usar frecuencias de cualquier par de *hits* correlativos para estimar las probabilidades condicionadas.

Hay que mencionar la analogía entre reglas directas y cadena de Markov de primer orden, ya que sus probabilidades condicionadas corresponden con la Confianza de reglas de secuencia directas de orden 2. Y así en adelante: se puede demostrar que un modelo de Markov de orden 2 es un modelo de secuencias directas de orden 3, etc.

La diferencia con el modelo de Markov es que éste es global y no local, lo que se refleja en que considera todas la páginas y no sólo aquellas con un Soporte amplio. El modelo de Markov es un modelo probabilístico, por lo que se podrán obtener inferencias.

Para construir el modelo de Markov (de primer orden y homogéneo) se reorganizan los datos en el *dataset* Page, ya descrito anteriormente, en que aparece el *referrer* y el *request*. Para un visitante cualquiera, el *request* en una observación debería ser el *referrer* de la siguiente. Según lo explicado en apartados anteriores se eliminaron los *referrer* en blanco (se explican por configuración en el navegador cliente). Además, hay aspectos técnicos como la recarga de páginas y la falta de páginas intermedias que el SAS Miner considera en su algoritmo. La matriz de transición tendrá en definitiva 57x57 probabilidades estimadas y no se publica en el presente estudio por razones de espacio.

Se empieza por evaluar en qué página un visitante tiene más posibilidades de empezar la visita. Para ello, consideramos las probabilidades de transición de “Inicio” mayores de un 1% en el siguiente gráfico:

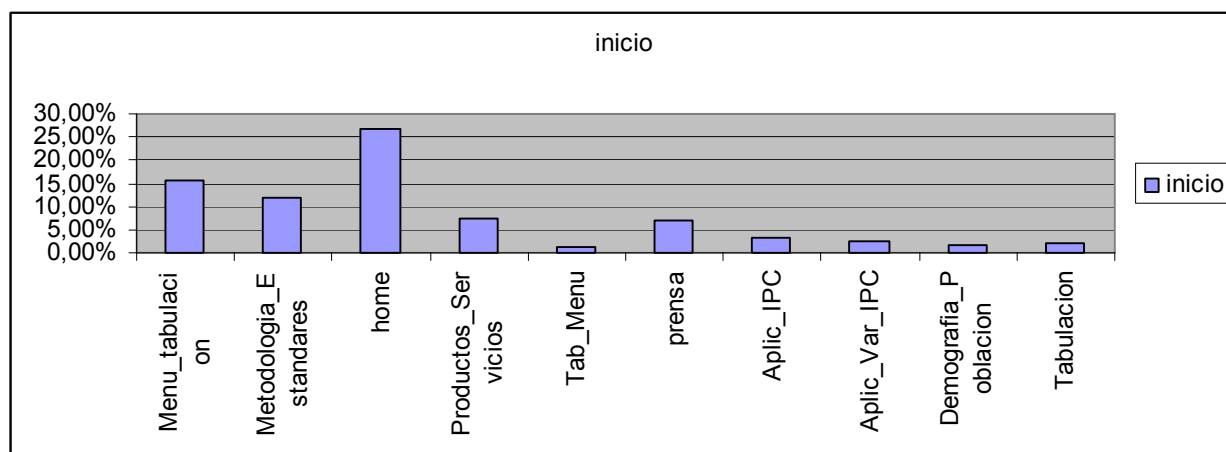


Gráfico 4.14. Gráfico con probabilidades de transición desde Inicio mayores de un 1%

La entrada más frecuente es a “home” con más de un 26% de los casos, seguidos por “Menu_tabulacion” con un 16% y “Metodologia_Estandares” con un 12%, lo que es consistente con la naturaleza de los visitantes que pertenecen a este *cluster*.

Las probabilidades de transición suman 1 por fila y columna, lo que no sucedía con las reglas de asociación. Ahora consideramos las páginas de salida más frecuentes, para lo que tomamos las probabilidades de transición de la columna Final, para claridad sólo las más significativas:

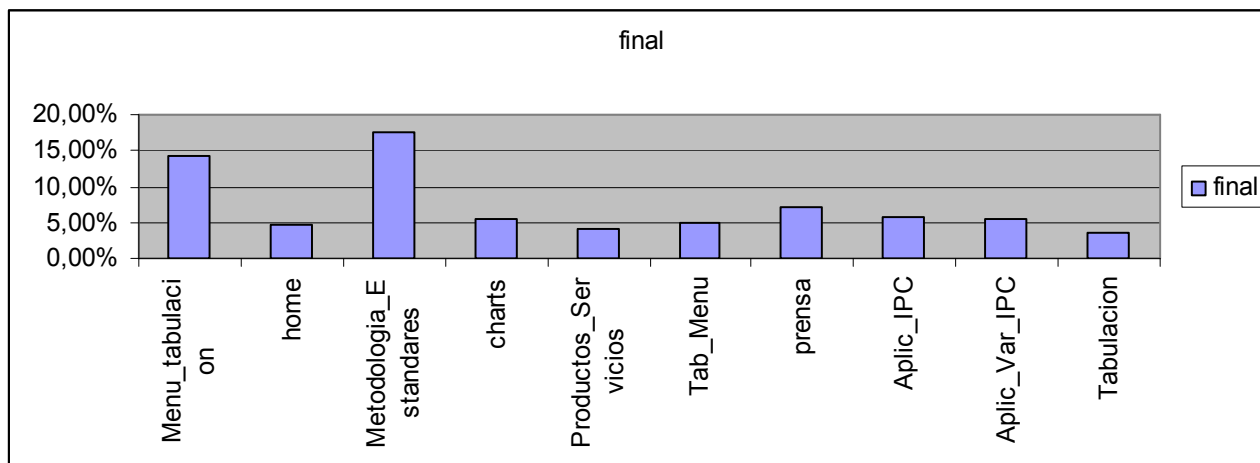


Gráfico 4.15. Gráfico con probabilidades de transición a Final mayores de un 1%

Se observa que es desde “Metodología_Estandares” donde más se abandona el *Website*, con un 17%, seguido por “Menu_tabulacion” con un 14% y “home”+“Charts” que conjuntamente (dada su naturaleza) suman un 10% de los casos.

Además, de la matriz de probabilidades de transición se puede obtener la ruta que conecta los nodos a través de las transiciones más probables. Desde “Inicio” conectaríamos con la página con la que tenga mayor probabilidad de transición que es “home” en el 26,7% de los casos. Desde ahí a “Charts” (72%) y de ahí de nuevo a “home” (posiblemente recarga), con lo que la ruta termina.

Se puede comparar con lo que sería la ruta anterior usando el índice de Confianza del algoritmo de las reglas de secuencia. Inicialmente, desde “Inicio” la Confianza más alta se alcanza igualmente en “home” con un 26,2% de los casos, seguido de “Charts” con un 64% de los casos y desde ahí de nuevo a “Charts” en un 30% de los casos. Esta información se representa gráficamente a continuación:

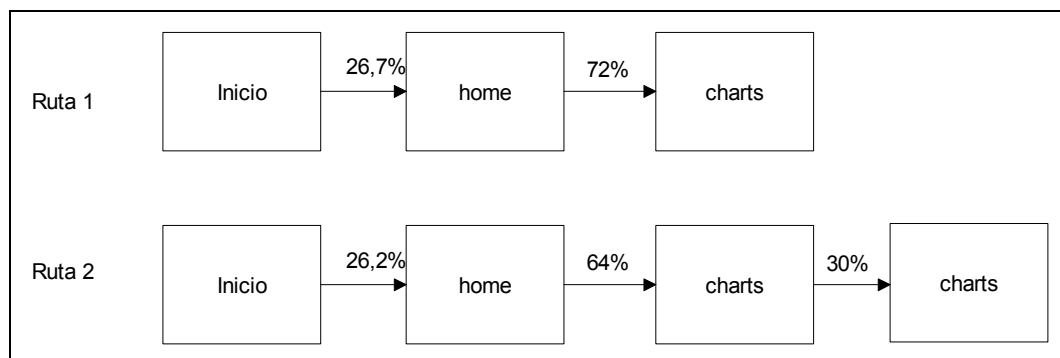


Figura 4.2. Ruta hacia delante de las transiciones más probables (Ruta 1 considera el modelo de cadena de Markov y Ruta 2 la de reglas de secuencia)

Las diferencias con la cadena de Markov son debidas a que las secuencias directas consideran sólo páginas que superan un cierto soporte.

Siguiendo la misma lógica, se puede construir una ruta hacia atrás como en la siguiente figura, que se inicia desde “Final”. La página con la mayor probabilidad de transición a “Final” es “Metodologia_Estandares” con un 17,5% y luego “Inicio” con un 50% de probabilidades de transición desde “Inicio” a “Metodologia_Estandares”. Por tanto, aquí la ruta se completa ya que llega desde “Final” a “Inicio”.

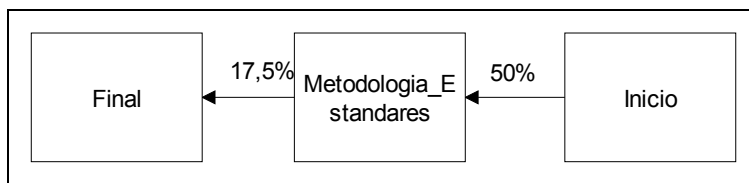
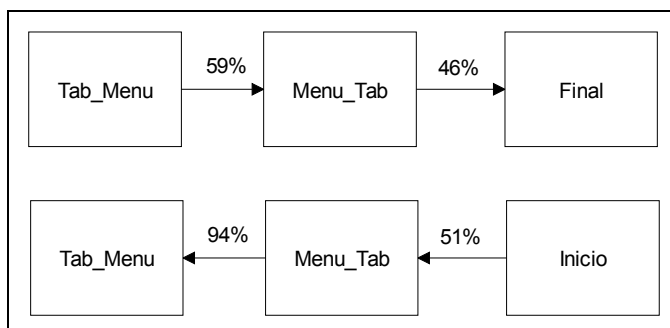


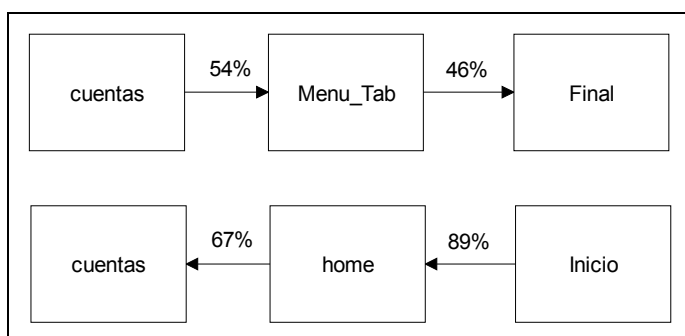
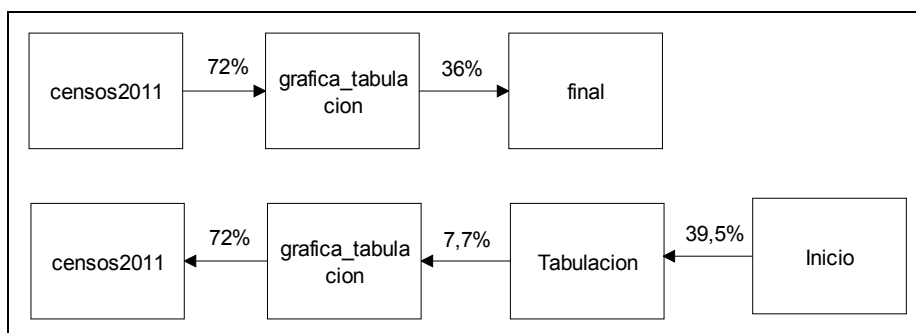
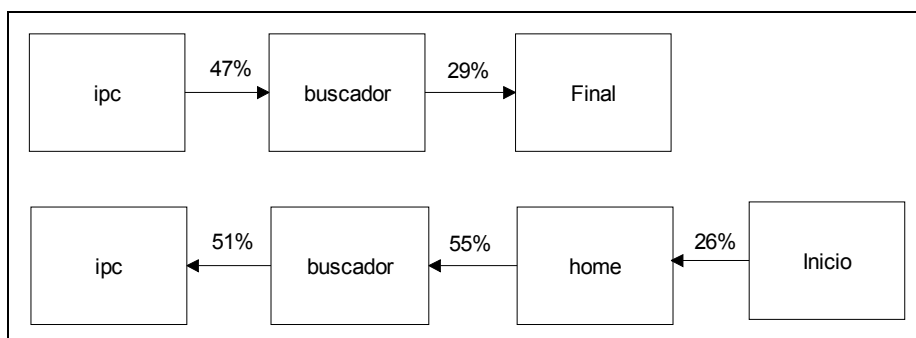
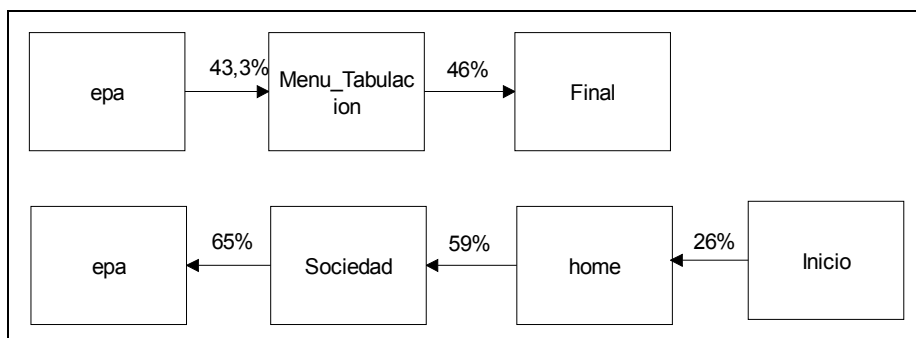
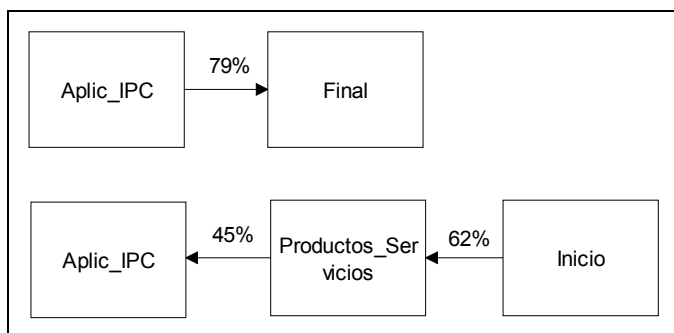
Figura 4.3. Ruta hacia atrás de las transiciones más probables

Estas dos rutas no son las que aparecerán con mayor seguridad. Para calcularlas habría que comparar todas las cadenas de Markov de todos los órdenes, lo que exigiría un excesivo poder computacional.

Alternativamente, las reglas de secuencia pueden ser la mejor opción, ya que aún siendo locales son fáciles de calcular e implementar, incluso para grandes *datasets*, si bien al no estar normalizado no sumarán 1.

Se concluye este análisis de cadenas de Markov describiendo las transiciones estimadas para algunas de las páginas más importantes:





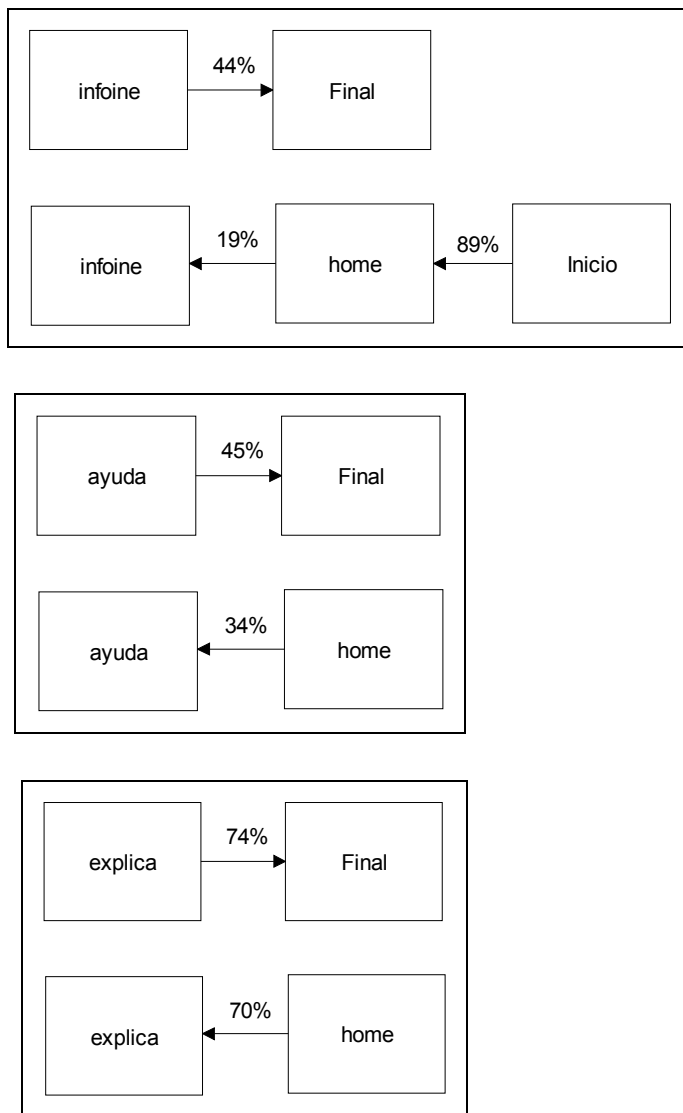


Figura 4.4. Transiciones estimadas para algunas de las páginas más importantes

Entre otros temas, llama la atención que desde “IPC” se salga de forma tan exhaustiva al “buscador” y desde ahí se salga del *Website*. Además entra mayoritariamente desde el buscador que a su vez viene de “home”, lo que indica la posibilidad de que no esté suficientemente bien presentada la opción al usuario.

Después de visitar “EPA” el usuario va mayormente a “Menu_tabulación”, lo mismo que “Cuentas”, pero a “Cuentas” se accede sobre todo desde “home”, a diferencia de “EPA”, al cual se accede desde Inbase (“Sociedad”)

La relación entre “censos2011” y “grafica_tabulación” viene motivada por la interrelación automática entre ambas páginas. Además, llama la atención que su inicio de sesión viene de una página exterior y no entra a través de “home”.

Los usuarios de “Explica” son especializados en el sentido de que después de ver la página de interés abandonan el Website. Por su parte, los usuarios que usan “infoine”, aunque vienen mayoritariamente de “home”, proceden de gran variedad de *sites* del INE, lo que indican que en alguna medida, antes de contactar, han navegado por el *Website*.

Otros resultados interesantes son que de “Tempus” se accede y procede mayoritariamente a/desde “prensa” (50% aproximado) y a la aplicación “meparezco” se procede desde “Censo2011” casi siempre y después se finaliza en la mitad de los casos.

Cabe destacar que desde “Tabulación” se accede al “buscador” como segunda opción más frecuente (20%), lo mismo que desde “intercensal” (10% de los casos). Además, a “Tabulacion” se procede desde el “buscador” en otro 20% de los casos y a “intercensal” desde “buscador” en otro 10%.

A “territoriales” se accede casi en un 90% de los casos desde “home” y en la mitad de los casos se sale del *Website* una vez se visita. Además a “geotempus” se accede como es natural en el 70% de los casos desde territoriales.

COMPARACIÓN DE LOS MODELOS

No es sencillo evaluar modelos locales, considerando además que en este estudio se propone evaluar conjuntamente modelos locales y globales. Para los modelos globales, como los *SEP* y las cadenas de Markov, la evaluación estadística se podría establecer mediante criterios computacionales (como el *cross-validation*, criterio Bootstrap, *bagging* de algoritmos genéticos, etc.), mediante criterios basados en funciones de pérdida (gráfico del Lift, curva ROC, etc.) o mediante funciones *score* (función AIC, BIC, etc.), entre otros.

Pero el problema es cómo compararlos con las reglas de secuencia. Para éstas y en ausencia de otras consideraciones, una simple función *score* para una regla de secuencia sería el Soporte, que daría la proporción de la población a la que la regla es aplicable. Alternativamente, si el objetivo es predecir el comportamiento de los visitantes en términos de probabilidad condicionada de moverse entre páginas debería usarse el Lift o la Confianza.

Entre los modelos considerados, **las secuencias directas y las cadenas de Markov se pueden comparar entre ellas** y pueden interpretarse fácilmente ya que tienen un significado único (las cadenas de Markov son reglas de secuencia directa con una probabilidad condicionada).

Las reglas indirectas se pueden interpretar como más descriptivas y más relacionadas con la asociación y derivan en el *Link Analysis* que se ha mostrado, que es una manera de representar la información de una forma fácilmente entendible.

Por otro lado, los SEP estudian el modelo desde otra perspectiva, en la que la dependencia del modelo es entre variables aleatorias más que entre ocurrencias.

Ahora se van a comparar de manera no exhaustiva los resultados obtenidos por las reglas directas con las de las cadenas de Markov. A partir de la Gráfico 4.8, con las reglas directas ordenadas por el Lift, se comprobaba que algunas de ellas eran realmente informativas, como en el caso de “Aplic_Var_IPC” \Rightarrow “Aplic_Var_IPC” y “Aplic_IPC” \Rightarrow “Aplic_IPC”, con Lifts por encima de 15. La siguiente destacable era la de “Productos_Servicios” \Rightarrow “Aplic_Var_IPC” con un Lift de 5,69, lo que nos indica que la probabilidad de visitar la Aplicación de Variación del IPC es casi 6 veces mayor si antes estabas en una página del apartado de “Productos y Servicios”.

Otras reglas no son informativas, por ejemplo la de “Menu_Tabulacion” \Rightarrow “Metodologia_Estandares”, con un Lift de poco más de 1.

Conclusiones similares se pueden sacar del modelo de la cadena de Markov, donde el Lift de cada transacción se puede calcular dividiendo las probabilidades de transición por las probabilidades del estado inicial.

Así, por ejemplo podríamos determinar que el Lift de “Menu_Tabulacion” \Rightarrow “Tab_Menu” sería de $0,3468 / 0,1127 = 3,08$, algo menor a las de reglas de secuencia (3,69), dado que estas últimas sólo consideran como candidatas de transición aquellas páginas cuyo Soporte es mayor que un umbral determinado, mientras que las cadenas de Markov consideran todo el *dataset*.

V.- SUMARIO DEL PROCEDIMIENTO Y RESÚMEN DE RESULTADOS

1. Justificación del estudio: El estudio de los Weblogs es un aspecto fundamental en el avance del conocimiento de lo que los usuarios de la Web demandan del INE, al ser el canal fundamental de interrelación con ellos.

El objetivo último del Organismo es proporcionar los datos estadísticos demandados y por tanto prevalece, además de la calidad del contenido, la oportunidad en la información proporcionada (que interese) y la calidad del soporte en que se entrega, que se ve tan influenciado por el constante cambio de tecnología. El apoyo en las nuevas tecnologías puede constituirse en una ventaja, haciendo más atractivo el acceso a la información estadística.

Una de las vías que se siguen para avanzar en ese camino es el análisis de la "caja negra" de su interrelación con el Organismo, es decir las peticiones al servidor Web, que es lo que se ha planteado en el presente estudio.

2. Objetivo: Este estudio tenía como finalidad la identificación de patrones de secuencia en el *Website* del INE, siendo un patrón una secuencia ordenada de páginas, posiblemente repetidas. La medida de importancia elegida determina los resultados del análisis, siendo las medidas más comunes las referidas a la probabilidad de que ocurra una secuencia determinada (Soporte) o la probabilidad condicionada de acceder a cierta página, una vez se han visto otras páginas antes (Confianza).
3. Organización de los datos: los datos se han extraído de un *logfile* que registra el acceso al *Website*. Se ha estructurado en ficheros transaccionales y planos. No se ha estructurado como matriz de datos para no perder el sentido temporal. El estudiado preprocesamiento de los ficheros, ha sido clave para la consecución de resultados válidos.
4. Análisis Exploratorio: Esta fase del análisis es necesaria para extraer conclusiones válidas. Habiendo tres variables continuas relativas a las características del visitante- el número de clicks por sesión, la hora y la duración de la conexión- se han eliminado los *outliers* y se ha "clusterizado" resultando varios comportamientos diferenciales, lo que nos ha permitido identificar grupos homogéneos de visitantes, concentrando el análisis en el principal.

En cuanto al perfil de visitantes en el *cluster* principal, se ha visto que existen distintos tipos en función del contenido al que accede. Se ha comprobado como interesante el contenido de las

búsquedas del usuario, de la que se podría inferir que considera al INE casi como proveedor universal de cualquier materia estadística imaginable. También se ha visto que una mayoría realiza pocos *hits* en su visita, en concreto más de la mitad realiza tres o menos clicks y más de una tercera parte abandona el *Website* en el primer hit. En otra aplicación de la técnica del *cluster*, se ha comprobado el comportamiento diferencial según las páginas que visita, el número de *hits* realizados, etc. Además, se ha realizado una comparativa con herramientas usadas actualmente para el análisis de logs, como el Google Analytics, razonando las diferencias y las posibles causas.

5. Especificación del modelo: El análisis de minería de datos en este caso es un ejemplo de modelo local. Se han comparado reglas de secuencia con el análisis de *links*. Posteriormente se han usado técnicas estadísticas más tradicionales basadas en todo el *dataset* con Sistemas SEP y cadenas de Markov.

Con estas últimas se ha extraído información interesante sobre las rutas más usuales que siguen los visitantes durante su visita. Por ejemplo, se ha visto que al subgrupo de “IPC” se entra sobre todo desde el “buscador”; que a “cuentas” se entra, a diferencia de “epa”, desde “home”; que a “tempus” se entra desde “prensa” y a “meparezco” desde “censo2011”.

Asimismo, con las *SEP* se han comprobado relaciones indirectas no identificadas con el algoritmo *a priori* del SAS Miner, como la relación muy significativa entre “epa”==>“IPC”; “Censos2011”==> “Aplic_Nombres”; “Aplic_Nombres”==> “Censos2011”; “Aplic_Nombres”==> “Tabulacion”.

Del nodo “Asociación” del Miner, también se extrae información interesante como las páginas de entrada (por ejemplo, “home” es inicio en el 88% de las ocasiones en que hay “home” en la visita; a las aplicaciones de IPC, a “Metodología_Estándares” y a “Información_INE” se entra muy a menudo desde un dominio externo al INE) y salida (por ejemplo, casi siempre se termina la visita después de algún “rss”; después de una visita a “Prensa”, casi dos terceras partes de los usuarios abandonan el *site* del INE) más frecuentes.

El *Link Analysis* nos da la representación gráfica de los resultados del nodo de “asociación”, mostrando la relación entre los nodos, la magnitud de dicha relación y el tamaño del nodo.

6. Comparación de modelos: Dada la dificultad de evaluación de modelos locales, también es complicado compararlos con modelos globales basados en modelos estadísticos, considerando

que están basados en presunciones distintas. Sin embargo, se ha visto que las reglas de secuencia directas se pueden comparar con las cadenas de Markov, poniendo de manifiesto que los resultados de los patrones más probables son asimilables.

7. Interpretación del modelo: las reglas se interpretan fácilmente, aunque exista el problema de tener que discernir entre su gran cantidad. Los modelos estadísticos globales, como las cadenas de Markov, hacen más fácil la selección de las reglas más relevantes, dado que son interpretables de manera inmediata.
8. Conclusiones: Como consecuencia del presente estudio, se abren varias posibilidades de investigación (por ejemplo, análisis de errores de *Weblogs*, adaptación de la Web y contenidos a los resultados de los análisis y seguimiento del cambio, análisis de las búsquedas en la Web del INE, etc.). El Sistema de Información propuesto es una vía válida para dicho análisis, considerando su actualización cada vez que se modifique la estructura del *Website* del INE y adaptándolo al objeto de estudio.

Las ventajas de dicho Sistema es su posibilidad de “customización” a los objetivos perseguidos además de tener el pleno control sobre el propio Sistema, lo que no ocurre con otras herramientas que tampoco permitirían ir más allá de análisis exploratorio de los datos. La desventaja principal es que bajo la arquitectura propuesta necesita, para grandes volúmenes de información (generalmente se analizarán uno o varios días de *logs*) una capacidad de proceso elevada. Una vía de mejora sería modificarla a otro entorno (nosql, entorno mainframe, etc.)

Los datos de secuencia, asociación y “clusterización” obtenidos pueden ser usados para mejorar la estructura del *Website*, para idear sistemas recomendadores, etc. y en todo caso, para aumentar el conocimiento del perfil de usuarios que visitan el *site*, objetivo inicial del estudio.

Otra vía de conocimiento del usuario vendría por una potenciación del uso (y análisis) de las redes sociales y de subscripciones y registros, a partir de los cuales se obtendría información socio-demográfica que poder utilizar en segmentaciones, *Customer Relationship Management* (CRM), sistemas personalizados, etc. El uso de la tecnología y esa capacidad de la Institución de ser referente en materia estadística en el ámbito nacional, podrían resultar herramientas de interés para atraer a los usuarios y potenciar la cantidad de accesos Web.

WEB CONTENT MINING

La minería de contenido, tiene como principal objetivo otorgar datos reales o finales a los usuarios que interactúan con la Web. Es decir, extraer información “útil” de los contenidos de las páginas web.

Generalmente la información disponible, se encuentra de forma no estructurada (minería de Texto), semi-estructurada y un poco más estructurada como es el caso de tablas html generadas automáticamente con información de bases de datos.

De acuerdo con Raymon Kosala y Hendrick Blockeel, la minería de contenido puede ser diferenciada desde dos puntos de vista; desde el punto de vista de la Recuperación de Información (IR) y desde el punto de vista de Base de Datos (DB).

Es decir, asistir en el proceso de recogida de información o mejorar la información encontrada por los usuarios, generalmente basada en las solicitudes hechas por ellos mismos (IR). Desde la perspectiva de DB principalmente se trata de modelar los datos e integrarlos en la Web a través de *queries* sofisticadas.

MINERÍA DE CONTENIDO DESDE EL PUNTO DE VISTA DE RECUPERACIÓN DE INFORMACIÓN

La recuperación de información es el proceso de encontrar el número apropiado de documentos relevantes de acuerdo a una búsqueda hecha en una colección de documentos. La IR y la *Web Mining* tienen diferentes objetivos, es decir, la *Web Mining* no busca remplazar este proceso. La *Web Mining* pretende ser utilizado para incrementar la precisión en la recuperación de información y mejorar la organización de los resultados extraídos.

La recuperación de información es muy utilizada en grandes empresas del mundo Web, las cuales hacen uso de este tipo de sistemas, las máquinas de búsqueda (google y altavista), directorios jerárquicos (yahoo) y otros tipos de agentes y de sistemas de filtrado colaborativos.

La diferencia principal, independientemente de las técnicas que usan, existente entre la Recuperación de la información y la Extracción de la Información recae principalmente en que uno recupera documentos relevantes de una colección y la otra recupera información relevante de dichos documentos. La IE se centra principalmente en la estructura o la representación de un

documento mientras que la IR mira al texto en un documento como una bolsa de palabras en desorden [Wilks Y. Information Extraction as a Core Language Technology Source Lecture Notes In Computer Science; Vol. 1299 Pages: 1 – 9, 1997]. Podemos decir que dichas técnicas son complementarias una de otra y usadas en combinación pueden generar valor añadido.

Datos no estructurados, semi-estructurados y estructurados, son los objetivos de la Extracción de Información, generalmente para los datos no estructurados se hace uso de técnicas de Lenguaje Natural. Dichas reglas son generalmente basadas en el uso de relaciones sintácticas entre palabras y clases semánticas. Reconocimiento de objetos de dominios tales como, nombres de personas y compañías, análisis sintáctico y etiquetado semántico, son algunos de los pasos para la extracción de información en documentos no estructurados.

Recientemente, se ha hecho uso de una tecnología llamada *Text Mining* , que hace referencia principalmente al proceso de extracción de información y conocimiento interesante, no trivial desde documentos no estructurados.

Las principales categorías de la tecnología Web *Text Mining* son *Text Categorization*, *Text Clustering*, *Association Analysis*, *Trend Prediction*.

Text Categorization: dada una predeterminada taxonomía, cada documento de una categoría es clasificada dentro de una clase adecuada o más de una.

Es más conveniente o sencillo realizar búsquedas especificando clases que buscando en documentos. Actualmente existen varios algoritmos de *Text Categorization*, dentro de los cuales encontramos, *K-nearest*, *Neighbor-algorithm* y *Naive Bayes Algorithm*.

Text Clustering: el objetivo de esta categoría es el de dividir una colección de documentos en un conjunto de clústeres tal que la similitud *intra-cluster* sea minimizada y la similitud *extra-cluster* sea maximizada. Podemos hacer uso de *Text Clustering* a los documentos que fueron extraídos por medio de una máquina de búsqueda. Las búsquedas de los usuarios referencian directamente a los clusters que son relevantes para su búsqueda. Existen dos tipos de *text clustering*, *clustering* jerárquico y *clustering* particional (G-HAC y k-means) [J. Wang, Y. Huang, G. Wu, and F. Zhang, Web Mining: Knowledge Discovery on the Web Proc. Int'l Conf. Systems, Man and Cybernetics (SMC '99), vol. 2, pp. 137-141].

MINERÍA DE CONTENIDO DESDE EL PUNTO DE VISTA DE BASE DE DATOS

La Web es una fuente enorme de documentos en línea que regularmente contienen datos semi-estructurados. La Extracción de información en la Web se afronta de diferente manera a lo antes hecho. Ahora hay que enfrentarse a un volumen extenso de documentos Web, a los documentos nuevos que aparecen con periodicidad y al cambio en el contenido de los documentos Web. Una gran parte de los documentos o páginas web contienen datos semi-estructurados y estructurados y generalmente o siempre contienen información a través de *Links* [Eikvil, L. "Information Extraction from World Wide Web - A Survey", Rapport Nr. 945, July, 1999. ISBN 82-539-0429-0].

El objetivo principal que tiene el *Web Content Mining* desde el punto de vista de BD es que busca representar los datos a través de grafos etiquetados. La publicación de datos semi-estructurados y estructurados en la Web ha crecido fuertemente en los últimos años y existe la tendencia a seguir creciendo. Sin embargo, el crecimiento ha sido preponderante en las "*hidden Web*" o páginas ocultas, las cuales son generadas automáticamente con datos de bases de datos a través de consultas hechas por usuarios.

Dichas páginas no son accesibles para los *crawlers* y para las máquinas de búsqueda no están a su alcance. Así pues, existe la necesidad de crear ciertas aplicaciones o herramientas para la extracción de información de tales páginas. Para la obtención de dicha información en las Web se hacen uso actualmente de los llamados "*wrappers*". Los *wrappers* pueden ser vistos como procedimientos para extracción de contenido de una fuente particular de información. La extracción de estos datos permite otorgar valor agregado a los servicios, por ejemplo, en los comparativos de compras, meta búsquedas, etc. Existen varios enfoques para la extracción de información estructurada; *manual wrapper*, *wrapper induction* y el enfoque automático.

El primero consiste en escribir un programa para extracción de información de acuerdo con los patrones observados en un *Website* en específico. Los segundos consisten en identificar un grupo de páginas de entrenamiento y un sistema de aprendizaje generará reglas a partir de ellas, finalmente dichas reglas serán aplicadas para obtener objetos identificados dentro de páginas web.

Finalmente, el método automático tiene como objetivo principal identificar patrones de las páginas web y luego usarlas para extraer información. Seguramente éste último es el método más utilizado en la actualidad para extraer información de la Web.

De acuerdo con WangBin [Wang, B. Liu, Z. "Web Mining Research," Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03), iccima, p. 84, 2003] las estructuras de Links permiten otorgar mayor información que otro documento normal. La *Web Structure Mining* se centra principalmente en la estructura de los *HiperLinks* de la Web, es decir, está interesada en la entrada y salida de *Links* de las páginas. Los *Links* que apuntan a una página pueden sugerir la popularidad de la misma, mientras que los *Links* que salen de la página demuestran los tópicos o la riqueza de contenido.

Algoritmos como el PageRank y los HITS son usados con frecuencia para modelar la topología de la Web. En PageRank, cada página web tiene una medida de prestigio que es independiente de cualquier necesidad de información o pregunta. En línea general, el prestigio de una página es proporcional a la suma de las páginas que se ligan a él.

PageRank es un valor numérico que representa lo importante que es una página en la Web. Para Google, cuando una página (A) enlaza a otra (B), es como si la página (A) que tiene el enlace, votara a la página enlazada (B). Mientras más votos tenga una página, más importante será la página. También, la importancia de la página que vota determina lo importante que es el voto. Google calcula la importancia de una página a partir de los votos que obtiene. En el cálculo del PageRank de una página se tiene en cuenta lo importante que es cada voto.

HITS (*HyperLink Induced Topic Research*) es un algoritmo interactivo que tiene como finalidad extraer el grafo de la Web para identificar "*Hubs*" y "*Authorities*". Entendemos como *Authorities* a las páginas que de acuerdo a un tópico son las que mejor posicionadas están. Los *Hubs* son aquellas páginas que ligan hacia las *Authorities*. El número y el peso de *Hubs* apuntando a una página determina el nivel de posicionamiento.

VII.- ANEXO SOBRE NOCIONES BÁSICAS DE TECNOLOGÍA WEB

Se definen a continuación algunas nociones y términos básicos dentro de la tecnología Web que son especialmente importantes en este estudio.

DIRECCIÓN IP

Es el dato que identifica de manera única cada equipo informático conectado a Internet o a cualquier otra red que utilice el protocolo TCP/IP. Suele representarse usando un formato de 4 números menores o iguales a 255 separados por puntos (por ejemplo: 194.179.36.23).

URL

Es el acrónimo de *Uniform Resource Locator* o "Localizador Uniforme de Recurso". Se trata de una cadena de caracteres con la cual se asigna una dirección a cada recurso (página, imagen, sonido, etc.) disponible en Internet. El formato general de una URL es: protocolo://dirección_máquina:puerto/directorio/fichero

En el caso concreto de la tecnología Web, el protocolo es HTTP y el puerto por defecto es el 80, de modo que las URL tienen la forma: http://dirección_de_la máquina/directorio/fichero

NAVEGADOR VS SERVIDOR WEB

La tecnología Web se basa en el protocolo HTTP y una arquitectura cliente/servidor, en la que la aplicación cliente suele ser conocida como navegador Web o *Web browser*, que es la utilizada por el usuario para acceder a los contenidos publicados por el servidor Web correspondiente.

Los navegadores Web más populares son: Microsoft Internet Explorer, Netscape, Opera, Firefox, Safari, etc.

Los servidores Web más utilizados en Internet son: Apache HTTP Server, Microsoft Internet Information Server, etc.

Muchas implementaciones de servidor Web permiten definir varios servidores virtuales sobre una misma instalación del servidor, de manera que cada uno de ellos puede tener una configuración y unos contenidos propios; desde el punto de vista del usuario, funcionan igual que si fueran servidores independientes.

SITIO WEB (O PORTAL O WEBSITE)

Se trata del conjunto de contenidos publicados en un servidor Web, relacionados lógicamente entre sí y que comparten una misma estructura: página inicial, menú de navegación, etc.; por ejemplo, el sitio Web de una empresa, el de un organismo oficial, etc.

En ocasiones, el término “portal” se utiliza para un tipo específico de sitio Web (el que agrupa contenidos de otros sitios) o para referirse únicamente a la página inicial, pero en este estudio lo consideraremos como equivalente a “sitio Web” o en inglés *Website*.

INTRANET VS WEB (PÚBLICA)

Ambos entornos utilizan la misma tecnología cliente/servidor, basada en el protocolo TCP/IP; más concretamente, suelen hacer referencia a la tecnología Web: protocolo HTTP y lenguaje HTML. La diferencia reside en su ámbito de utilización: una Web pública está accesible desde Internet y está a disposición de cualquier usuario desde cualquier equipo conectado a la red. Por otro lado, una intranet está aislada, teniendo limitado su acceso desde la red particular/privada de una empresa, organización, etc., y está dirigida a sus empleados o miembros. En ambos casos puede existir o no autenticación de usuarios (usualmente mediante un nombre y una contraseña).

Una *Extranet* es un híbrido de ambos entornos y suele referirse a un acceso identificado, a través de Internet, a la intranet de una entidad. No se contempla en este estudio.

CONTENIDOS ESTÁTICOS VS CONTENIDOS DINÁMICOS

Los servidores Web funcionan básicamente como servidores de ficheros: reciben una petición de lectura de un determinado fichero HTML y el servidor lo devuelve (junto con sus imágenes y otros ficheros asociados) sin ningún tratamiento adicional. Todavía muchos servidores funcionan únicamente de ese modo; a esos contenidos se los denomina estáticos.

Sin embargo, existe la posibilidad de que los contenidos devueltos por el servidor no sean únicamente los ficheros almacenados, sino que pueden ser variables o dinámicos (accediendo, por ejemplo, a bases de datos) en función de determinadas circunstancias (parámetros de entrada; usuario), lo que da una flexibilidad mucho mayor a los servidores Web. La implementación de estos servidores dinámicos ha ido evolucionando y se ha realizado utilizando diferentes tecnologías, entre otras las siguientes:

1. CGI (*Common Gateway Interface*), en la que el servidor Web ejecuta un programa escrito en un lenguaje de programación común (C, Perl, etc.) cuya salida estándar es devuelta directamente al navegador; dicho de otro modo, el programa "devuelve HTML".
2. Lenguajes de *scripting*; que consisten en incorporar a los ficheros HTML unas marcas especiales que son interpretadas y ejecutadas por el servidor Web. Algunas de estas tecnologías son JSP, PHP, ASP, cada una con su propio lenguaje de programación.
3. Java *servlets*, que funcionan de manera parecida a los CGI pero utilizan un lenguaje y arquitectura concretos (Java), integrados con el propio servidor Web y con funcionalidades adicionales de seguridad, acceso a bases de datos, etc.
4. Servidores Domino (Servidores Lotus Notes/Domino). Etc.

LOGS (TRAZAS)

Se trata de la grabación o registro de una actividad. Un *log* de servidor Web es el registro de la actividad que dicho servidor deja normalmente en una serie de ficheros de texto, en los que se almacenan las peticiones realizadas (fecha, hora, URL, dirección IP del cliente, ...) y el resultado (código de retorno/error). Los formatos más habituales de estos *logs* de servidor Web, son Common Logfile Format y Extended Log File Format, que se describen a continuación.

Los formatos Common Logfile Format y Extended Log File Format son similares; el segundo, definido como extensión del primero, está soportado por la práctica totalidad de servidores Web existentes.

Dirección IP del cliente

La dirección IP desde la que se ha accedido al servidor. En ocasiones, esta dirección puede ser la de una máquina que actúa como intermediario (*proxy*, *firewall*), por lo que el servidor Web no almacena la dirección IP real del cliente, dificultando la identificación de usuarios únicos, seguimiento de sesiones, etc.

Dirección del servidor

El nombre del servidor utilizado para acceder al recurso. Este campo es útil sobre todo si sobre una misma instalación de servidor se están ejecutando diferentes servidores virtuales, ya que permite separar los accesos a cada servidor virtual para su tratamiento.

Nombre de usuario

Si el usuario se ha autenticado previamente en el servidor Web, se almacena su nombre. En otro caso, se deja vacío este dato.

Fecha y hora

Fecha y hora del servidor en el momento que se realiza la petición.

Request (o Petición o Recurso)

La petición realizada por el cliente. Este dato suele incluir el método HTTP utilizado (GET, POST, etc.), el recurso (página, imagen, etc.) solicitado, y la versión del protocolo HTTP utilizada.

Código de respuesta (status)

El código HTTP de respuesta que identifica el resultado de la petición.

Tamaño de la respuesta

El tamaño, normalmente en bytes, del mensaje del servidor enviado como respuesta a la petición.

Referrer (remitente)

Dato opcional enviado en el encabezado HTTP; contiene la URL del recurso origen desde el que se realizó la petición. Por ejemplo, en el caso de solicitar una página a través de un enlace, este dato contendría la URL de la página que incluye el enlace; en la petición de una imagen, contendría la URL de la página que contiene dicha imagen. Es una información útil a la hora de seguir la secuencia de páginas visitadas por un usuario.

User agent

Una cadena de texto opcional enviada por el navegador en el encabezado de la petición HTTP que identifica la plataforma (versión de navegador, de sistema operativo, etc.) del cliente. En la práctica este dato no tiene un formato normalizado, y puede ser falseado fácilmente, por lo que su tratamiento se realiza mediante técnicas heurísticas que no son totalmente fiables.

TERMINOLOGÍA DE ESTADÍSTICAS WEB

En el tratamiento estadístico del uso de sitios Web se utilizan diferentes términos que conviene definir de manera precisa, ya que a menudo se confunden y se utilizan con diferentes significados.

Request (Petición o hit o acceso)

Una petición consiste en la solicitud de un fichero al servidor Web, que recibirá un hit o acceso y lo registrará en una entrada (línea) del fichero de *log* al mismo tiempo que enviará un código de respuesta junto con el fichero (si está disponible). En la práctica estos tres términos son equivalentes y tienen mayor utilidad en un ámbito técnico.

Página vista/visitada

Una página web, que es solicitada y percibida por los usuarios como una unidad, está formada generalmente por un conjunto de ficheros (HTML, imágenes, hojas de estilos, etc.); por tanto, cada página vista (o visitada) por el usuario suele generar varios hits en el servidor. La “página visitada” es la unidad básica para conocer el grado de popularidad de los contenidos de un sitio Web.

Visita, sesión de usuario

Una visita (o sesión de usuario) está formada por el conjunto de páginas accedidas por un usuario durante una misma sesión de trabajo; generalmente se considera que la sesión de trabajo se mantiene mientras el tiempo entre la vista de dos páginas consecutivas no supere un determinado umbral o *timeout* no siempre fácil de determinar (en este estudio se ha considerado 30 minutos).

El estudio de las sesiones de usuario proporciona información referida al comportamiento de los usuarios, en contraposición a los datos anteriores que están referidos al funcionamiento del servidor.

Usuario (único)

En principio, un usuario único se corresponde con cada persona física que accede al servidor en el período de tiempo estudiado. Cuando no existe autenticación en el servidor (habitualmente realizada mediante un nombre y una contraseña), la identificación de usuarios reales se suele basar en la dirección IP, aunque dicha correspondencia no es totalmente fiable.

VIII.- ANEXO DE BIBLIOGRAFIA

- [1] Baeza-Yates, R. Castillo, C. Marin, M. and Rodríguez, A. "Crawling a Country: Better Strategies than BreadthFirst for Web Page Ordering", WWW Conference/Industrial Track, ACM, pp. 864-872. Chiba, Japan, 2005.
- [2] Baeza-Yates, R. Excavando la Web. El profesional de la información. v13, n1, 2004
- [3] Liu, B. and Chen-Chuan Chang, Kevin. Editorial: "Special Issue on Web Content Mining". WWW 2005 Tutorial, Page 1-4, 2005.
- [4] Scotto, M. Sillitti, A. Succi, G. Vernazza, T. "Managing Web-Based Information", International Conference on Enterprise Information Systems (ICEIS 2004), Porto, Portugal, April 2004. Page 1-3
- [5] Etzioni O. "The World Wide Web: quagmire or gold mine?" , In Communications of the ACM 39(11). 1996
- [6] www.wikipedia.org
- [7] Henzinger M. "Web Information Retrieval an Algorithmic Perspective", European Symposium on Algorithms, p 1-3, 2000
- [8] Baeza-Yates, R. Pobrete, B. "Una herramienta de minería de consultas para el diseño del contenido y la estructura de un sitio Web" Actas del III Taller Nacional de Minería de Datos y Aprendizaje TAMIDA2005", pp.39-48, 2005
- [9] Kosala, R. and Blockeel, H. Web Mining Research: A Survey. ACM SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining. Page 1-9, 2000.
- [10] Galeas, P. Web Mining by Patricio Galeas. <http://www.galeas.de/WebMining.html>
- [11] http://www2.ing.puc.cl/gescopp/Sergio_Maturana/SAG/WebMining.html